UNIVERSITY OF ALBERTA

TRANSCRIPTIONAL MAPPING IN HUMAN CHROMOSOME 22Q11.2

BY

POLLY BRINKMAN-MILLS ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Science.

In

Molecular Biology and Genetics

Department of Biological Sciences

Edmonton, Alberta
Fall, 1999

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-47011-3

Canada

UNIVERSITY OF ALBERTA

Library Release Form

Name of Author: Polly Brinkman-Mills

Title of Thesis: Transciptional Mapping in Human Chromosome 22q11.2

Degree: Master of Science

Year this Degree Granted: 1999

Permission is hereby granted to the University of Alberta to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, except as hereinbefore provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

P. Lly Brinkman-Mills

Polly Brinkman-Mills
9611-73 Avenue
Edmonton, Alberta
T6E 1B3

DATE: October 1 1999

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled "Transcriptional mapping in human chromosome 22q11.2" submitted by Polly Brinkman-Mills in partial fulfillment of the requirements for the degree of Master of Science in Molecular Biology and Genetics.

Heather McDermid

Dr. H.E. McDermid (supervisor)

Dr. D.B. Pilgrim

Dr. M.A. Walter

DATE: Sept 28/99

ABSTRACT

Human chromosome 22 is predicted to be gene rich, as well as being the location of numerous syndromes including cat eye syndrome (CES). CES is a developmental disorder characterized by ocular coloboma, anal atresia, preauricular malformations, heart and kidney defects, and characteristic facial features. Individuals with CES typically have a supernumerary, isodicentric, and bisatellited chromosome [idic(22)pter-q11.2], resulting in four copies of this region. The minimal duplicated region is defined as the pericentromeric 2Mb of 22q.

The purpose of this study was to identify genes within the CES critical region (CESCR) using two gene identification techniques, exon trapping and genomic sequence analysis. Using these techniques, two putative genes were identified. One gene, CES86, shows a muscle specific transcript and illustrates an interesting genomic organization. Another gene, CES38, was not predicted by genomic sequence analysis, stressing the need for greater than one gene identification technique to be employed for any given region.

ACKNOWLEDGMENTS

Heather McDermid

Thanks to Heather for being an excellent supervisor and for always providing sound judgement and encouragement throughout my project.

Drs. Mike Walter and Dave Pilgrim

Thanks to the members of my supervisory committee for providing invaluable suggestions.

Dr. Ali Riazi

Thank-you Ali, for giving me a piece of your project.

Drs. Valerie Trichet and Andrew Wong

Thank-you for providing me with RNA samples and the expertise to use them.

Angela Johnson and Dana Shkolny

Thanks guys, for teaching me a lot of what I know.

Dr. Bruce Roe

Thanks to our collaborator at the University of Oklahoma, who provides us with excellent genomic sequence.

Graham Banting and Dr. Song Hu

Thanks for all the help and troubleshooting with the computers !

Stephanie Maier

You're a great friend, thanks for helping to keep me sane. Thanks sooooo much for so many things !

The McDermid lab (past and present)

Thanks to everyone for being awesome co-workers.

My parents

Thank-you for raising me with such high standards and always encouraging me to reach for all my goals.

My Husband, Jeffrey

Thank-you sweetie for always being my shoulder to cry on when nothing seemed to work. Thank-you for encouraging me, for always

being there no matter how late I worked, and for still loving a busy wife.

TABLE OF CONTENTS

Chapter 1: Introduction **Page**

Human Chromosome 22 1

Chromosome 22q is Thought to be Gene-Rich 1

Anomalies Associated With Chromosome 22 2

Cat Eye Syndrome 4

- Features 4
- History and Etiology of CES 7
- Other Duplications Associated with CES 8
- Definition of the Minimal CES Critical Region (CESCR) 9
- Molecular Pathology 12
  - Ligand / Receptor or Signal Transduction Molecules .. 13
  - Structural Proteins 14
  - Transcriptional Regulation 14

The Human Genome Project 16

Chromosome 22: Genetic & Physical Mapping and Cloning 17

Cloning of the CESCR 19

Chromosome 22 Sequencing Consortium 20

Transcriptional Mapping 20

- Gene Identification Techniques 21
  - CpG Island Identification and Cloning 21
  - Exon Trapping 23
  - cDNA Selection 24
  - Comparative Mapping 25
  - Computer Analysis of Genomic Sequence 27

22q Pericentromere: Rich in Repetitive DNA and Unprocessed 29
Pseudogenes and Posing a Challenge to Transcript Mapping
in the CESCR
Research Objectives 35

Chapter 2: Materials and Methods

Exon Trapping 44
Preparation of Bacterial or P1-derived Artificial Chromosome
(BAC or PAC) DNA 44
Preparation of Plasmid DNA 45
Southern Blotting and Hybridization of DNA probes 46
Northern Blots and Hybridization 48
Cycle Sequencing 49
Screening Human cDNA Libraries 50
Polymerase Chain Reaction 51
Reverse Transcription PCR 52
Subcloning 53
Rapid Amplification of cDNA ends (RACE; 3' and 5') 54
Database Searching and Genomic Sequence Analysis 55

Chapter 3: Results

A) Characterization of Putative Exons Within the CESCR 67

Exon Amplification 67
Colony PCR 68
Sequencing and BLAST Searches 69
Mapping Within the CESCR 70

Determination of Loci Number 71
Screening cDNA Libraries 73
B) Characterization of Putative Genes in the CESCR 74
1) Partial Cloning and Characterization of a Putative 74
Gene, CES86
Mapping cDNA Clones Within the CESCR 74
Sequence Analysis of Three cDNAs: Ex86 K1a, ... 75
hEST1 and hEST2
Expression Analysis of CES86 77
5' End Cloning of hEST1 78
3' End Cloning of hEST1 79
Southern Analysis of CES86 80
CES86 Polymorphism Can Be Used to Determine 82
Parent of Origin of CES Chromosomes
Genomic Sequence Analysis of PAC 143I13 83
2) Putative Genes Flanking CES86 84
A) An Insect-Derived Growth Factor-Like Gene 84
(IDGFL)
B) A Phosphatidyl Synthase-Like Gene (PSL) 85
C) A Putative Single Exon Gene, BTPUTR 86
D) The Interleukin-17 Receptor Gene 87
Genomic Sequence Analysis of PAC 109L3 88
3) Partial Cloning and Characterization of a Putative 89
Gene, CES38
Mapping the Ex 38 cDNA Clone Within the CESCR ... 89
Sequence Analysis of the Ex38 Q1 cDNA 90
Expression Analysis of the Ex 38 Q1 cDNA 90
Southern Analysis of CES38 91
Genomic Sequence Analysis of PAC 238M15 92

Chapter 4: Discussion

Identification of Genes in the CESCR 129

1) Amplification of Putative Gene fragments 130

Exon Trapping Summary 130

Effectiveness of Exon Trapping 131

Other Possible gene Identification Techniques 133

Drawbacks of Exon Trapping Within the CESCR 134

2) Identification of a Putative Gene, CES86 135

Future Experiments 139

3) Identification of a Putative Gene, CES38 141

Future Experiments 141

4) Partial Characterization of a Putative Gene, BTPUTR 143

5) Partial Characterization of a Putative gene, PSL 144

Developmental Defects in CES, a Molecular Etiology 144

Future Research 145

Identification of Genes From the CESCR 145

Further Characterization of CES38 and Other Candidate Genes 145

Conclusions 147

References 150

LIST OF FIGURES

Number	Title	Page
Figure 1-1	Ideograms of chromosome 22	36
Figure 1-2	A cat eye chromosome	37
Figure 1-3	The cat eye syndrome critical region (CESCR)	39
Figure 1-4	A partial physical map of the CESCR	40
Figure 1-5	The chromosome 22 sequencing consortium	41
Figure 1-6	Human-mouse comparative map	42
Figure 2-1	Exon trapping system	60
Figure 2-2	A detailed flow chart of Marathon cDNA amplification protocol	62
Figure 2-3	Genomic sequence analysis flow chart	64
Figure 3-1	Colony PCR products	93
Figure 3-2	A partial physical map of the CESCR	95
Figure 3-3	The Ex86 K1 cDNA	96
Figure 3-4	Three Ex 86 cDNAs	98
Figure 3-5	The sequence of the Ex86 K1a cDNA	100
Figure 3-6	The sequence of the hEST1 cDNA	101
Figure 3-7	Expression analysis of CES86	103
Figure 3-8	Further Expression analysis of CES86	104
Figure 3-9	RT-PCR analysis of CES86	105
Figure 3-10	Ex86 partial hybrid panel	106
Figure 3-11	CES86 identifies a polymorphic *Pst* site	107
Figure 3-12	Restriction analysis of CES86 genomic region	109
Figure 3-13	Determination of parent of origin using CES86	110
Figure 3-14	Annotated sequence of PAC143I13	112
Figure 3-15	Expression analysis of hEST52444 (PSL)	113

Figure 3-16 Expression analysis of hEST 46414 (BTPUTR) 115
Figure 3-17 Expression analysis of IL-17 receptor gene 116
Figure 3-18 Annotated sequence of PAC 109L3 118
Figure 3-19 The Ex 38 Q1 cDNA 119
Figure 3-20 The sequence of the Ex38 Q1 cDNA 121
Figure 3-21 Expression analysis of CES38 122
Figure 3-22 CES38 partial hybrid panel 123
Figure 3-23 Annotated sequence of PAC 238M15 125
Figure 4-1 Exon trapping versus genomic sequence analysis .. 149

LIST OF TABLES

Number	Title	Page
Table 1-1	Frequency of CES congenital defects	 43
Table 2-1	A list of the cDNA libraries used	 65
Table 2-2	A list of all the primers used	 66
Table 3-1	Summary of exon trapping results	 126
Table 3-2	A list of the putative exon sequences and an interpretation of the BLASTN and TBLASTX search results	 127
Table 3-3	A list of the putative exons and the number of loci they likely represent in the genome	 128

ABBREVIATIONS

ALD: adrenoleukodystrophy gene

ASD: atrial septal defect

BAC: bacterial artificial chromosome

BLAST: basic local alignment search tool

BTPUTR: big three prime UTR

CBP: CREB binding protein

cen: centromere

CES: cat eye syndrome

CESCR: cat eye syndrome critical region

CHD: congenital heart defect

cM: centimorgan

CML: Chronic Myelogenous leukemia

CMT1A: Charcot-Marie-Tooth disease type 1A

CTAFS: conotruncal anomaly face syndrome

dbEST: EST database

DGS: DiGeorge syndrome

ES: Ewing sarcoma

EST: expressed sequence tag

FISH: florescent in situ hybridization

GCPS: Grieg cephalopolysyndactyly

GGT: γ-glutamyl transferase gene

GSP: gene-specific primer

HGP: human genome project

HNPP: hereditary neuropathy with liability to pressure palsies

htgs: high throughput genomic sequence

HUGO: human genome organization

IDGFL: insect-derived growth factor-like

IGKVP3: immunoglobulin kappa light-chain pseudogene

IL-17R: interlukin-17 receptor gene

IRP: island rescue PCR

kb: kilobase pairs

KCNMB3L: calcium-activated potassium channel gene three-like

Mb: megabase

mya: million years ago

NF1: neurofibromatosis type 1

nr: non redundant

OMIM: online Mendelian Inheritance in Man
(http://www.ncbi.nlm.nih.gov/Omim/)

ORF: open reading frame

PAC: P1 artificial chromosome

PAF: platelet activating factor

PMP22: peripheral myelin protein 22

PSL: phosphatidyl synthase like

RACE: rapid amplification of cDNA ends

RFLP: restriction fragment length polymorphism

RTK: receptor tyrosine kinase

RT-PCR: reverse transcription PCR

RTS: Rubenstein-Taybi syndrome

STS: sequence tagged site

TAPVR: total anomalous pulmonary venous return

tel: telomere

TOF: Tetrology of fallot

UTR: untranslated region

UTS: untranslated sequence

VCFS: velocardiofacial syndrome

VSD: ventricular septal defect

vWF: vonWillebrand factor gene

YAC: yeast artificial chromosome

CHAPTER 1: INTRODUCTION

Human Chromosome 22

Human chromosome 22 is the second smallest autosome with an estimated size of 53 Megabases (Mb) (http://www.sanger.ac.uk/HGP/Chr22/) which accounts for approximately 1.7 to 1.9 % of the total human genome (Buetow et al. 1994; Morton 1991). Low resolution G-banding (400 band level) divides the q arm of chromosome 22 into three regions: two G-light bands, q11 and q13, separated by a G-dark band, q12. High resolution G-banding (850 band stage) further subdivides the q arm into 10 subbands as shown in Figure 1-1. Characteristic of the acrocentrics (13, 14, 15, 21, and 22) are relatively small p-arms that share similar structure and organization. These p-arms consist mostly of heterochromatin and repetitive DNA sequences. The only known active genes on the acrocentric p-arms are the ribosomal RNA genes which are located in the secondary constriction, or stalk, at 22p12 (Schmickel & Knoller 1977; Kaplan et al.1987).

Chromosome 22q is Thought to be Gene-Rich

Chromosome 22 is thought to be gene rich based on its banding pattern. G-light bands are generally early replicating and contain a high proportion of unmethylated GC-rich sequences (Cuny et al. 1981; Sentis & Fernandez-Piqueras 1993), identifying large proportions of euchromatin, while G-dark bands are late replicating and contain a high proportion of AT-rich sequences, signifying a heterochromatic composition. G-light bands also contain the highest density of CpG islands (Bernardi 1995) which are short dispersed regions of unmethylated DNA with a high frequency of CpG dinucleotides relative to the bulk genome. Outside the

islands, CpG is depleted to approximately 20 % of the expected value and 60 to 90 % are methylated on the C ring (Bird 1986). They are useful as landmarks for genes since about 60 % of genes are associated with a CpG island.

The mainly G-light (and therefore gene-rich) composition of 22q, in addition to its small size and the large number of chromosomal rearrangements associated with many syndromes and cancers, make it the focus of much research. Genetic and physical mapping of chromosome 22 has been underway for many years and more recently intensive sequencing efforts have begun. The sequence of 22q will be invaluable for identification of genes associated with various diseases and syndromes.

Anomalies Associated With Chromosome 22

Chromosome 22 is associated with numerous chromosomal anomalies of both acquired and congenital origin. Acquired conditions resulting from 22q rearrangements include three well characterized cancer-related translocations: Chronic myelogenous leukemia (CML; [t(9;22)(q34;q11)]) (Rowley 1973), Burkitt's lymphoma ([t(8;22)(q24;q11)]) (Berger et al. 1979; Haluska et al. 1987), and Ewing sarcoma (ES; [t(11;22)(q24;q12)]) (Aurias et al. 1984; DeLattre et al. 1992; May et al. 1993).

In addition to the cancer-related rearrangements, 22q is also the site for numerous congenital conditions. Some of these are due to partial duplications of 22q and some are associated with partial deletions. Among these are a group of syndromes associated with an interstitial microdeletion in 22q11.2. These include DiGeorge Syndrome (DGS; OMIM # 188400; OMIM, 1999; McKusick, 1998), velocardiofacial syndrome

(VCFS; OMIM # 192430), conotruncal anomaly face syndrome (CTAFS; Burn et al. 1993) and familial and sporadic isolated conotruncal heart defects (Goldmuntz et al. 1993).

In addition to the 22q deletion syndromes mentioned, there are also two conditions associated with the presence of a duplication of part of 22q. These include cat eye syndrome (CES), which is the result of a duplication of the proximal region of 22q11, and the recurrent t(11;22) derivative 22 syndrome [der(22)t(11;22)(q23;q11)]. The t(11;22) translocation is the only recurrent non-Robertsonian congenital translocation known in humans (Fraccaro et al. 1980; Zackai & Emanuel 1980). The only viable unbalanced offspring, resulting from a 3:1 meiotic segregation, will carry a der(22) chromosome and will have a duplication of 11q23 to 11qter and 22pter to 22q11, including the region duplicated in CES. The phenotype of individuals carrying this der(22) does show some overlap with the phenotype of CES (Fraccaro et al. 1980) however it is difficult to determine which features are a result of duplication of proximal 22q and which are the result of duplication of distal 11q.

Besides the congenital and cancer-related conditions associated with chromosome 22q, it is also the site for a host of other syndromes including: Opitz G syndrome type II (OMIM# 145410), Spino-cerebellar ataxia-10 (OMIM# 603516), Wardenburg-shah syndrome (OMIM# 277580) and 22q13.3 deletion syndrome (Nesslinger 1994; Wong 1998).

Cat Eye Syndrome

- **Features**

Cat eye syndrome (CES) is a rare developmental disorder with an estimated incidence of approximately one in 50,000 to 150,000 (OMIM# 115470). It is characterized by the association of a variety of congenital defects summarized in Table 1-1 (Modified from Mears 1995). The phenotype of CES is highly variable, even within a single family, in both the features present and their severity (Schinzel et al. 1981). CES derives its name from the iris coloboma, even though this feature is present in only about half of CES individuals. A typical coloboma of the iris is the result of failure of closure of the optic fissure during the sixth week of embryonic development (Moore & Persaud 1993). Coloboma can be unilateral or bilateral, and in addition to the iris, can also affect the choroid and/or the retina.

In 1878 Haab first noted an association between coloboma and imperforate anus, which was later called CES by Gerald (Gerald et al. 1968). Imperforate anus, or membranous anal atresia, results from failure of the anal membrane to rupture at the end of the eighth week of development (Moore & Persuad 1993).

The frequency of anal atresia is second only to preauricular malformations, which is the most common feature in individuals with CES. Preauricular malformations include skin tags (or appendages) and pits (or sinuses). Approximately five weeks into development, the auricle of the external ear begins forming from six swellings called auricular hillocks, which develop around the margins of the first branchial (pharyngeal) groove in the embryo. Preauricular tags are the result of development of accessory auricular hillocks. Most preauricular pits are

the remnants of the first branchial groove (Moore & Persaud 1993). Besides preauricular malformations other typical facial features associated with CES include: hypertelorism (wide-set eyes), down-slanting eyes, micrognathia (small jaw), and low set and/or malformed posteriorly rotated ears.

Also associated with CES are congenital heart defects (CHDs). Development of the heart begins in the third week in the cardiogenic area located rostrally in the embryo (Moore & Persaud 1993). A pair of endothelial strands called angioblastic cords appear and soon become canalized to form endothelial heart tubes. These tubes approach each other and become fused to form a single heart tube. By the end of the third week the endothelial heart tube has linked up with blood vessels in the embryo to form a primitive cardiovascular system. Also by the end of the third week the heart begins to beat. The cardiovascular system is the first organ system to reach a functional state. The primitive heart tube will become partitioned into four chambers between the fourth and seventh weeks. The most common CHDs associated with CES are Tetrology of Fallot (TOF) and total anomalous pulmonary venous return (TAPVR). TOF is a combination of four cardiac defects consisting of:

1) Pulmonary Stenosis

   - narrowing of the pulmonary trunk/artery leaving the right ventricle

2) Ventricular Septal Defect (VSD)

   - Failure of the membranous part of the interventricular septum to develop, or incomplete closure of the interventricular foramen, leaving a hole between the left and right ventricles. This allows mixing of oxygenated and deoxygenated blood, resulting in the reduction in oxygenation of the blood circulated to the body.

3) Overriding Aorta

- the aorta lies over the VSD and receives blood from both ventricles (both oxygenated and deoxygenated)

4) Hypertrophy of the Right Ventricle

- a larger than normal right ventricle

(Moore & Persaud, 1993)

TAPVR (MIM 106700) occurs when the pulmonary veins carrying oxygenated blood from the lungs fail to connect with the left atrium, and instead connect directly to the right atrium or one of its venous tributaries. This defect is not detrimental for the fetus due to high pulmonary vascular resistance and shunting of blood through the foramen ovale (connection between the atria). However, at birth the pulmonary vascular resistance drops and increased blood flow to the right heart and lungs results in progressive congestive heart failure and pulmonary arterial hypertension (Correa-Villasenor et al. 1991). If not surgically corrected, TAPVR has a high mortality rate in the first year of life (http://www-medlib.med.utah.edu/reprogen/research/tapvr/index.html). Theories regarding the development of the pulmonary veins are controversial. It is not clear whether the common pulmonary vein forms as an outpouching of the left atrium or whether it is formed at the lungs and then subsequently "directed" to the left atrium.

Other organ-systems may also be affected in individuals with CES. Renal defects are seen in approximately 45 % of patients and most often include kidney agenesis (an absent kidney) and renal hypoplasia (underdeveloped kidneys). The permanent kidneys (metanephroi) develop during the fifth to eighth weeks from the ureteric buds (metanephric diverticuli) and the metanephric mesoderm. The ureteric bud is a dorsal outgrowth from the mesonephric duct that grows into a mass of metanephric mesoderm (Moore & Persaud). Unilateral renal

agenesis occurs when the ureteric bud fails to develop or with early degeneration of this ureteric bud. It is a relatively common occurrence, seen about once in every 1000 newborn infants (Moore & Persaud). Anomalies associated with the skeletal and genital systems have also been seen, but usually at a lower frequency.

Mental retardation is also a feature of CES. Individuals with CES are typically reported to have mild to moderate mental retardation, although mental development in the normal range is also seen.

Considering the developmental processes of the affected organs in CES, it is likely that CES is a result of abnormal gene expression during the first nine weeks of embryological development.

- **History and Etiology of CES**

In 1965, Schachenmann et al. reported the presence of a small supernumerary chromosome in patients with ocular coloboma and anal atresia. Cytogenetic analysis revealed the presence of satellites on both ends of these supernumerary chromosomes (Schachenmann et al. 1965; Toomey et al. 1977), suggesting an acrocentric chromosomal origin (chromosome 13, 14, 15, 21 or 22). Due to its small size, chromosome banding studies could not clearly demonstrate the origin of the "cat eye chromosome" (Buhler et al. 1972) and chromosomes 13 (Krmpotic et al. 1971), 14 (Pfeiffer et al. 1970) or 22 (Buhler et al. 1972; Schinzel et al. 1981) were implicated. A chromosome 21 origin was not favoured due to the lack of phenotypic overlap between CES and Down's Syndrome. Schinzel et al. (1981) generally believed in a chromosome 22 origin because of the partial phenotypic overlap of CES with trisomy 22 and partial trisomy associated with the der(22) of the 11;22 translocation.

Definitive proof was not provided until 1986 when McDermid et al. demonstrated by quantitative dosage analysis and in situ hybridization that the 22q11.2-specific probe p22/34 (locus D22S9) was present in four copies in six cat eye syndrome patients carrying the cat eye chromosome. The typical cat eye chromosome is now known to be an isodicentric bisatellited chromosome derived from an inverted duplication of the short arm and proximal long arm of chromosome 22 [idic(22)pter-22q11.2] (McDermid et al. 1986). McDermid et al. call this the "CES chromosome".

- **Other Duplications Associated with CES**

Although the dicentric bisatellited CES chromosome is the most common form of duplication seen in patients with CES, there are individuals with CES carrying other forms of duplications. Reiss et al. (1985) and Knoll et al. (1995) each reported individuals (patients LW and SK respectively) carrying an interstitial duplication of the proximal region of chromosome 22q. Such individuals carry three copies of the minimal duplicated region known to cause CES. Patient SK presented with preauricular pits, TAPVR, hypertelorism, down-slanting palpebral fissures, congenital hearing loss, absent right kidney and testicle, and moderate motor delay (Knoll et al. 1995). LW had colobomata, preauricular pits, hypertelorism, down-slanting palpebral fissures and developmental delay (Reiss et al. 1985).

Besides the presence of an interstitial duplication, three cases of CES have also been reported due to the presence of a supernumerary ring chromosome 22 [r(22)] (El-Shanti et al. 1993; Ohashi et al. 1993; Frizzley et al. 1999). Ring chromosomes are formed when a chromosome undergoes two breakages (forming two acentric fragments) and the broken ends of the chromosome fuse into a ring structure (Thompson et

al. 1991). Ring chromosomes may experience instability at mitosis, when the sister chromatids attempt separation at anaphase. There may be breakage of the ring resulting in mosaicism for various fragments including larger or smaller rings, double rings, or ring fragments.

In 1995, Mears et al. reported the characterization of a r(22) chromosome in three generations. The presence of the r(22) in the grandfather (CM13) and father (CM14) was associated with a normal phenotype. The proband (CM15) presented with all the cardinal features of CES including coloboma, preauricular pits and tag, micrognathia, cleft palate, undescended testes, imperforate anus, TAPVR, interrupted aortic arch, VSD, ASD, polycystic kidneys, and urethral reflux. This patient died at 17 days of age and therefore his developmental potential is unknown. By quantitative dosage and RFLP analysis, and fluorescence in-situ hybridization (FISH), Mears (1995) found that the grandfather and father contained three copies of the loci D22S9 and D22S43, whereas the proband contained four copies of these loci. This suggested that the r(22) in the proband had undergone doubling, likely due to its instability in the previous generations. All three individuals harboured the r(22) in approximately 95 % of their lymphoblasts.

- **Definition of the Minimal CES Critical Region (CESCR)**

The most common form of duplication associated with CES is a supernumerary, dicentric, and bisatellited chromosome, containing two copies of 22pter→ q11.2. Mears et al. (1994) and McTaggart et al. (1998) have found that the duplication breakpoints are clustered in two intervals within 22q11.2, and have therefore classified CES chromosomes into two types based on the location of the two breakpoints required to generate them. The smaller type I CES chromosomes are symmetrical, with both

breakpoints located within the proximal interval (between D22S427 and D22S36, ~ 3.3 Mb from the centromere; see Figure 1-2). The larger type II CES chromosomes can be either symmetrical, with both breakpoints located in the distal interval (between CRKL and D22S112, ~ 6.3 Mb from the centromere), or asymmetrical, with one breakpoint located in each of the two intervals.

Although the supernumerary idic(22) is the most common duplication associated with CES, characterization of patients such as SK and CM15, who carry other forms of duplication, allow definition of the critical region for CES. The critical region for CES (CESCR) is the smallest duplicated region of proximal 22q11.2 which results in the CES phenotype. Definition of the current CESCR shown in Figure 1-2 was accomplished by characterization of duplications of two previously mentioned patients.

The distal boundary of the CESCR was delineated by Mears et al. (1995) using the patient (CM15) with a supernumerary double r(22) chromosome. The breakpoint of this patients r(22) chromosome was found to be between the locus D22S57 and the gene ATP6E (See Figure 1-2). This narrowed the CESCR to the region between the centromere and the locus D22S57, a region of approximately 2 Mb (McDermid et al. 1996). The finding of patients with an interstitial duplication in proximal 22q11.2 exclude the short arm of chromosome 22 from the critical region.

The proximal boundary of the CESCR was provisionally defined by Mears (1995) by characterization of an interstitial duplication in patient SK (Knoll et al. 1995). The proximal breakpoint of this patient's interstitial duplication was found to be between the loci D22S795 and D22S543, a distance of approximately 1 Mb from the centromere, dividing the CESCR into proximal and distal halves. It must be noted however that this patient did not have coloboma or imperforate anus, two cardinal features of CES.

It is not possible to determine if the absence of these two features is due to the highly variable nature of this syndrome, or if the genes causing these two features map centromeric of locus D22S795. It is also possible that the overexpression of a gene or genes within the interstitial duplication but distal to the CESCR could have modified SK's phenotype. However, comparing the duplication overlap between the r(22) patient and SK, the region of focus is narrowed to the distal CESCR, a region of approximately 1 Mb. As well, the proximal CESCR is known to be rich in low copy and interspersed repetitive DNA as well as duplicated, non-processed, truncated gene fragments (McDermid, personal communication; S. Minoshima,1998) and therefore is likely gene-poor.

The only gene reported to date in the CESCR is the ε subunit of vacuolar H^+ - ATPase (Baud et al. 1994). ATP6E is located approximately 2 Mb from the centromere, just proximal to the locus D22S57. In addition to ATP6E, the gene for death agonist BID maps just distal to the locus D22S57 (Footz et al. 1998). Since BID maps outside the duplication region defined by the r(22) patient (CM15) it cannot be responsible for the cardinal features of CES. This child died at 17 days of age and therefore mental and physical development could not be assessed. Therefore it is possible that overexpression of BID could contribute to abnormal development of these features.

- **Molecular Pathology**

The goal of molecular pathology is to explain why a given genetic change should result in a particular clinical phenotype. In order to understand the clinical manifestations associated with CES we must be able to explain how duplication of the CESCR affects the function(s) of the gene product(s) involved. As we move from cataloging genes in the CESCR to understanding their function, we must speculate on the types of gene products which could cause the phenotype associated with CES. Because CES is a duplication syndrome, it's effects must be due to the extra copies of a gene or genes in the CESCR. When a gene(s) present in more or less than the normal number of copies produces a phenotypic effect, it is described as being "dosage sensitive". Dosage effects may be seen when a gene product interacts with other proteins or DNA sequences (Fisher & Scambler 1994). In most cases, what matters is not the correct absolute level of a gene product, but the correct relative levels of various interacting proteins (Strachan & Read 1996). Because of the interaction, phenotypic effects caused by dosage sensitive gene products may be modified by changes elsewhere in the genome. As well, overexpression of certain gene products may increase the sensitivity of the embryo to environmental factors such as infection or teratogens. This may result in variable phenotypic expression, even within a single family, which is indeed the case with CES.

We can speculate that the products of dosage sensitive genes may belong to a group of inherently dosage sensitive functions including: intercellular interactions (recognition, adhesion, communication), ligand/receptor signaling systems, signal transduction signaling systems, morphogens, structural gene products which co-operate with each other in interactions with a fixed ratio, and gene products which are

involved in transcriptional regulation (Fisher & Scambler 1994). A number of dosage sensitive genes involved in the production of phenotypic effects of various syndromes have been identified and will be discussed.

- **Ligand / Receptor or Signal Transduction Molecules**

A cell surface receptor and its ligand as well as intracellular transduction molecules are gene products which might be expected to be dosage sensitive. A ligand molecule frequently transmits its message to a cell via a cell surface receptor which is often coupled to a complex of intracellular transduction molecules. The proper interactions and functions of these proteins may require a strict stoichiometric ratio, and increased or decreased levels of any one of the molecules may result in a clinical phenotype. For example, deletions or mutations of the RET gene, which codes for a receptor tyrosine kinase (RTK), has been implicated in Hirschprung disease (MIM 142623). RTKs are cell surface molecules that transduce signals for cell growth and differentiation and are homologous to β-subunits of heterotrimeric G-proteins (Romeo et al. 1994; Edery et al. 1994). Improper proportions of β and γ subunits of G-proteins can disturb the formation of the normal protein complex and disrupt its normal function.

Deletions of 17p13.3, including the LIS-1 gene, are associated with Miller-Dieker lissencephaly. This disease is characterized by a brain malformation manifested by a smooth cerebral surface and abnormal neuronal migration (Dobyns et al. 1991; Reiner et al. 1993). The LIS-1 gene encodes a subunit of the heterotrimeric brain platelet-activating factor (PAF) acetylhydrolase (Hattori et al. 1994). PAF and PAF

acetylhydrolase are likely important components in the formation of the brain cortex during differentiation and development (Hattori et al. 1994).

- **Structural Proteins**

Many intracellular structural proteins interact with each other in a fixed ratio. Hemizygosity for the structural molecule elastin is responsible for the supravalvular aortic stenosis of Williams syndrome (MIM 194050; Ewart et al. 1993). It is likely that decreased gene dosage of elastin results in an insufficient quantity of protein products for the correct assembly of oligomers during a sensitive stage of development of the heart.

Both overexpression and underexpression of peripheral myelin protein 22 (PMP22) is associated with a specific neuropathy. Duplication of a 1.5 Mb region of 17p11.2 results in increased gene dosage of PMP22 which causes Charcot-Marie-Tooth disease type 1A (CMT1A; MIM 118220; Lupski et al. 1992; Patel et al. 1992). Deletion of the same region, resulting in underexpression of PMP22, causes hereditary neuropathy with liability to pressure palsies (HNPP; MIM 162500; Chance et al. 1993). The phenotypic effects caused by altered expression of PMP22 are thought to be due to an abnormal balance in the myelin of peripheral nerves (Strachan & Read 1996).

- **Transcriptional Regulation**

Increased or decreased dosage of many transcription factors may prove particularly likely to result in dosage effects because transcription factors often participate in competition for promoter sites and in the

assembly of multimeric complexes, where even a relatively subtle perturbation of one protein leads to altered stoichiometry with other subunits. In humans, haploinsufficiency of GLI-3, a member of the GLI-*Kruppel* gene family, causes Grieg cephalopolysyndactyly syndrome (GCPS; MIM 175700; Vortkamp et al. 1991, 1992). GCPS is characterized by polysyndactyly of the hands and feet, macroephaly, a broad base of the nose with mild hypertelorism and a prominent forehead. In *Drosophila*, a C2H2 zinc-finger protein called Kruppel, can form homodimers which repress transcription when the concentration of Kruppel is high. However, at low concentrations the monomer appears to act as a transcriptional activator, using the same DNA target sequences as are recognized by the repressor (Sauer & Jackle 1993).

In humans, Rubinstein-Taybi syndrome (RTS; MIM 180849), characterized by broad thumbs and great toes, characteristic facies, mental retardation, pulmonary stenosis, keloid formation in surgical scars, large foramen magnum, and vertebral and sternal anomalies (Berry 1987; Rubinstein & Taybi 1963), is caused by the loss of one copy of the CBP gene (Petrij et al. 1995). CBP is a CREB (cyclic adenosine 3',5'-monophosphate response element binding protein) binding protein, a nuclear protein participating as a co-activator in cyclic-AMP regulated gene expression (MIM 600140). The decreased dosage of CBP underlies the developmental abnormalities seen in RTS.

In humans, deletions or mutations of the PAX6 (MIM106210) gene cause aniridia (MIM 106200), which may be characterized by cataracts, glaucoma, corneal pannus, nystagmus, and foveal hypoplasia. Hemizygosity of the orthologous Pax6 gene in mouse results in the "small eye" phenotype (Ton et al. 1991). Schedl et al. (1996) have generated yeast artificial chromosome transgenic mice carrying the human PAX6 gene. Mice carrying multiple copies of the PAX6 show severe eye

abnormalities due to overexpression of this gene. In humans, overexpression of PAX6, due to a duplication of chromosome 11p12→13, results in eye abnormalities (Aalfs et al. 1997). The PAX6 gene product possesses a paired domain, a homeodomain, a serine/threonine-rich carboxy-terminal domain, and structural motifs characteristic of certain transcription factors (Ton et al. 1991). The PAX6 transcription factor likely regulates the transcription of a variety of tissue-specific genes.

The Human Genome Project

The Human Genome Project is an international project whose ultimate goal is to obtain a complete description of the human genome by DNA sequencing. This international collaboration was begun in the mid-1980s in the U.S.. Following the U.S. lead, national human genome programs have also been established in many countries including Canada, Japan, and throughout Europe. The Human Genome Organization (HUGO) was established in 1988 to coordinate the different national efforts (McKusick, 1989).

The complete nucleotide sequence of the human genome is only one of several goals of the HGP. Until present, the major emphasis of the HGP was the construction of high resolution genetic and physical maps. Sequence-ready physical maps composed of PAC, BAC, cosmid and/or fosmid clones preceded the ultimate physical map, the complete human genome sequence. The initial goal of the HGP was to complete the human genome sequence by 2005. To some extent, innovative sequencing technologies has allowed and will continue to allow, faster and cheaper automated sequencing. At present, the estimated date for completion of the complete human genome is the end of 2003 (Collins

1998; Goodman 1998). In the course of completing the sequence, a "working draft" achieving at least 90 % coverage, will be produced by the end of 2001. Other goals of the HGP include: sequencing technology development; human sequence variation studies; functional genomics technology development; comparative genomics (complete sequence of *C. elegans*, Drosophila, and mouse genomes) and identification of other model organisms; analysis of ethical, legal and social implications; development of more and better bioinformatics and computational studies; and training of skilled genomics researchers (Collins 1998; Goodman 1998; The Sanger Centre 1998).

Chromosome 22: Genetic & Physical Mapping and Cloning

Following with the ultimate goal of the HGP, genetic and physical mapping of human chromosome 22 has been underway for many years. In 1991 Dumanski et al. constructed one of the first detailed linkage maps of the entire q arm of chromosome 22. They placed 22 loci, defined by 30 polymorphic markers, on their map, representing a genetic distance of 110 cM. In 1994 Buetow et al. further refined the genetic maps, estimating chromosome 22 as 98 cM. Although genetic mapping has been a useful tool for ordering probes along the chromosome, it does have some limitations. In order to be informative, the polymorphic markers used must show a high percentage of heterozygosity. As well, genetic maps show relatively low resolution, since they rely on meiotic recombination events.

In order to produce a denser and more accurate map of chromosome 22, amenable to sequencing, other techniques were utilized. Physical mapping began with the construction of chromosome 22-specific somatic cell hybrid panels (Budarf et al. 1996). Somatic cell hybrids are cell lines that are typically constructed by fusing human cells

and rodent cells (usually mouse or hamster). The hybrid cells are initially unstable; most of the human chromosomes fail to replicate in subsequent rounds of cell division, and are lost (Strachan & Read 1996). This gives rise eventually to various hybrid cell lines containing a full set of rodent chromosomes plus one to a few human chromosomes or fragments of chromosomes. Sub-chromosomal mapping is possible using hybrid cell lines containing fragments of specific chromosomes. The human sub-chromosomal fragments may result from naturally occurring translocations or deletions or they may be artificially induced rearrangements. Budarf et al (1996) produced a somatic cell hybrid panel of chromosome 22 assigning over 300 markers to 24 unique regions or "bins". Although the order of markers within these regions cannot be determined using this technique, their use of markers from various sources facilitated the integration of physical and genetic maps. The markers used by Budarf et al. (1996) were also used by Bell et al. (1995) to anchor contigs of overlapping yeast artificial chromosome (YAC) clones spanning large regions of chromosome 22. The most complete YAC map of chromosome 22 has been constructed by Collins et al. (1995). They were able to place 620 markers on 705 YACs with an average probe density of one per 67 kb.

YAC clones are capable of carrying inserts > 1000 kb and have provided an efficient means to develop a physical contig of a region integrating a variety of markers. However, YAC clones are not ideal for direct genome sequencing because the YAC cannot be easily separated from the yeast genomic DNA and they also are often plagued with chimerism, rearrangements, and deletions (Kim et al.1996; McDermid et al. 1996). To circumvent these problems, YAC contigs are converted to maps based on other types of large insert vectors. The bacterial artificial

chromosome (BAC) and P1 artificial chromosome (PAC) are two vectors which can carry inserts ranging in size from 100-300 kb. These clones are stable and can be easily manipulated and directly used for genome sequencing. Sequence-ready physical maps are constructed by screening for bacterial artificial clones (BACs) or P1-artificial clones (PACs) using a high density of STSs, ESTs and/or known genes. A contiguous array of genomic clones is assembled by landmark (STSs, ESTs, genes, end clones etc.) mapping. Kim et al. (1996) have assembled a BAC-based physical map of approximately 80 % of chromosome 22q consisting of >400 markers over >600 BAC clones.

Cloning of the CESCR

A long-range physical map of the region duplicated in the typical CES chromosome, spanning a distance of approximately 3.6 Mb from the centromere to D22S36, was constructed by McDermid et al. (1996) using pulsed field gel electrophoresis (PFGE). This allowed the ordering of probes on the map and the estimation of actual distances (in kb) between probes. As well, McDermid et al. (1996) have assembled a YAC contig containing about half of the region between the centromere and D22S181. A number of these YACs were rearranged and/or deleted, and two gaps were present in the contig. Subsequently, a BAC/PAC based map was assembled by Johnson et al. (1999) spanning a distance of 1.5 Mb from D22S543 to D22S181. This PAC/BAC map has been the basis for transcriptional mapping within the CESCR and sequencing of many clones is currently underway.

Chromosome 22 Sequencing Consortium

Chromosome 22 is the second smallest chromosome, is comparatively gene rich, and harbours numerous rearrangements associated with many syndromes and cancers. This chromosome has therefore been the focus of intense sequencing, and thus is likely to be the first chromosome to be sequenced in its entirety. Sequencing of chromosome 22 is a collaborative effort of four major sequencing centers:

1) The Sanger Centre (http://www.sanger.ac.uk/HGP/Chr22/)

2) The Advanced Center for Genome Technology at the University of Oklahoma (http://www.genome.ou.edu/human.html)

3) Keio University School of Medicine (http://131.113.190.2/seqpub/keio.html)

4) The St. Louis Genome Sequencing Center (http://genome.wustl.edu/gsc/index.shtml)

Figure 1-4 illustrates chromosome 22 and the corresponding sequencing centers responsibilities for regional sequencing. BAC/PAC clones within the CESCR are being sequenced at the University of Oklahoma under the direction of Dr. Bruce Roe.

Transcriptional Mapping

Transcriptional mapping is the process of identifying genes in cloned DNA. Once a chromosomal region has been physically mapped and a contig of overlapping clones has been assembled, transcript mapping is performed. Completion of the transcript map for the human genome will be facilitated by the sequence of the total genomic DNA,

projected to be completed by 2003 (Collins 1998). In the meantime, attempts to produce transcript maps for many chromosomal regions are underway. Most of the efforts are concentrated on regions involved in human diseases and syndromes, including chromosome 22q (Riazi 1998; Wong 1998; Gong 1996; Minoshima et al. personal communication). Transcriptional mapping in regions that have genomic sequence available is simplified by the aid of computer analysis (gene and exon prediction programs). Identifying genes in unsequenced chromosomal regions is much more time consuming and relies on a variety of commonly used methods including:

- CpG island identification and cloning
- Exon trapping
- cDNA selection
- Comparative mapping

- **Gene Identification Techniques**

- **CpG Island Identification and Cloning**

CpG islands are short dispersed regions of unmethylated DNA with a high frequency of CpG dinucleotides relative to the bulk genome. About 60 % of genes are associated with a CpG island. Nearly all housekeeping or widely expressed genes are expected to have a CpG island, which is usually found at the 5' end of the gene usually covering the transcription start site, the promoter region and part of the coding sequence (Bird 1986, et al. 1987; Gardiner-Garden & Frommer 1987). Approximately 40 % of tissue-specific or limited expression genes also

contain a CpG island (Larsen et al. 1992). Unlike housekeeping genes, these CpG islands are not biased toward the 5' end and are often found in the body of the gene or at the 3' end (Toniolo et al. 1984; Gardiner-Garden & Frommer 1987). The average length of a CpG island is approximately 1 kb.

The relevance of the colocalization of CpG islands with predominately the 5' ends of many genes is not well understood. However, the presence of a CpG island can, by itself, give rise to relatively open and active chromatin (Kundu & Rao 1999). Tazi and Bird (1990) have demonstrated an alternative chromatin structure at CpG islands. They reported hyperacetylation of histones H3 and H4 and a deficiency of histone H1 in CpG island nucleosomes as well as the presence of nucleosome-free regions, all features of transcriptionally active chromatin.

CpG islands are useful as landmarks for genes because a large proportion of genes are associated with islands (Bird 1986). The high frequency of CpG dinucleotides within islands generates sites for rare-cutting restriction enzymes such as *Eag* I, *Not* I, *Sac* II and *Bss*HII which are methylation sensitive. Such enzymes are expected to cut much more frequently in CpG islands than in the bulk genome because of the likelihood of a sequence recognition site occurring within an island, as well as the site being unmethylated here. This technique has allowed the identification of many CpG islands and genes including: 28 CpG islands in Xq24-28 (Tribioli et al. 1992), 13 transcripts and 7 genes in the 4p16.3 region (Carlock et al. 1992; John et al. 1994), the cystic fibrosis gene (Rommens et al. 1989) and 12 novel genes in the human major histocompatibility complex class III region (Sargent et al. 1989).

As well, Valdes et al. (1994) has developed a PCR based approach for isolating transcribed sequences adjacent to CpG islands. Island rescue PCR (IRP) allows amplification of the region between a rare-cutter

restriction site within a CpG island and a nearby Alu repetitive element. Alu repeats are expected to occur approximately once every four kb in the human genome (Strachan & Read 1996). IRP has allowed the identification of cDNA clones for the neurofibromatosis type 1 (NF1) gene as well as nine other genes from chromosome locations 4p16.3 and 17q21 (Valdes et al. 1994).

Advantages of this technique are that it is a technically simple approach, it is expression-independent and it requires analyzing a smaller number of clones than other gene identification techniques. Disadvantages of CpG island identification include: the inability to identify genes without CpG islands, the identification of CpG islands associated with pseudogenes, and the high level of background resulting from CpG dinucleotides not associated with any detectable genes (John 1994).

- **Exon Trapping**

Exon trapping (or amplification) was developed by Buckler et al. (1991) as a method to rapidly and efficiently isolate exon sequences from cloned genomic DNA by virtue of selection for functional 5' and 3' splice sites. Pools of genomic DNA (from PAC, BAC, or cosmid) to be screened for exons are subcloned into an intron which is present within a specialized mammalian expression vector (pSPL1, pSPL3 or pSPL3B). The pSPL derivatives are propagated in *E. coli* , and the DNA is then isolated and transfected into COS cells. COS cells are derived from African green monkey CV-1 cells, containing the integration of a segment of the SV40 (Simian virus 40) genome with a defective origin of SV40 replication (Strachan & Read 1996). The integrated SV40 segment in COS cells allows any circular DNA which contains a functional SV40 origin of replication to replicate independently of the cellular DNA. Transcription

occurs in the COS cells from the SV40 promoter present in the pSPL vector and the RNA undergoes splicing under control of the host cell's RNA splicing machinery. When the cloned DNA fragment contains a recognizable exon in the proper orientation, spicing can occur between the vector and insert sequences. The splicing donor and acceptor sites in the vector are provided by two exons with an intron from the HIV *tat* gene. After transfection, total or cytoplasmic RNA is screened by RT-PCR for the acquisition of an exon from the genomic fragment. A unique fragment is produced when splicing has occurred between spice sites from the subcloned genomic fragment and pSPL. A diagram summarizing the exon trapping procedures can be found in Figure 2-1 in the materials and methods section.

The exon trapping methods and vector have been used to isolate and identify a variety of genes such as the Huntington's disease gene (HD collaborative Research Group 1993) and the neurofibromatosis type 2 (NF2) gene (Trofatter et al. 1993). The biggest advantage of this technique is that it is expression independent. Disadvantages of the exon trapping system include: the inability to identify single or two exon genes, the amplified fragments are small (usually < 200 bp) and are difficult to use in further characterization, the possible identification of unprocessed pseudogenes, and the possible identification of non-coding sequences which coincidentally have correct splicing signals (cryptic splice sites).

- **cDNA Selection**

cDNA selection (or direct selection) was developed to identify cDNAs encoded by large genomic regions (Lovett et al. 1991; Parimoo et al. 1991). The protocols are based on the hybridization of cDNA fragments which have been blocked for repetitive DNA elements (either an entire

library of cDNAs or cDNAs directly produced from a tissue(s)) to an immobilized genomic clone (YAC, PAC, BAC, cosmid etc.). Nonspecifically bound cDNAs are removed and the clone-specific cDNAs are eluted and amplified. This technique has been widely used for the identification of numerous genes in the genome including the BRCA1 gene located at 17q21 (Miki et al. 1994).

Advantages of the cDNA selection technique are that direct capturing of cDNAs may eliminate the need for obtaining full-length cDNAs (versus exon trapping), and single exon genes and genes without CpG islands can be identified. The main disadvantage of this system is that it is expression dependent, therefore it may be hard to find cDNAs that are tissue or time specifically expressed or at very low abundance. As well, this technique may identify cDNAs which hybridize to pseudogenes that have not deviated extensively from the active gene.

- **Comparative Mapping**

Since all organisms are related through a common evolutionary tree, the study of one organism can provide valuable information about others. Comparisons between distantly related genomes (e.g.: human and *Drosophila*, *S. cerevisiae* or *C. elegans*) may provide insight into the universality of biological mechanisms and provide experimental models for studying complex processes. Comparisons between more closely related genomes (e.g.: human and mouse) may provide insight into the details of gene structure and function. The power of comparative mapping arises from the fact that there is considerable selection pressure to conserve biologically important sequences such as coding sequences and regulatory regions. In contrast, less significant sequences and non-

coding DNA usually accumulate mutations comparatively rapidly and are not well conserved between species.

The mouse is currently the best mammalian model organism for studying human genetic disease and syndromes. Human and mouse share synteny over small to moderate sized subchromosomal regions. This refers to two orthologous genes that are syntenic in two or more species. "Conserved linkage" refers to conservation of both synteny and gene order of orthologous genes between species. Most genes in the human genome are expected to have a mouse orthologue. The sequence similarity of coding DNA between human and mouse is usually about 70-90 % while the respective polypeptide sequences show a slightly higher degree of sequence similarity, often within the 80-95 % range (Strachan & Read 1996). This degree of similarity between the human and mouse genomes is very useful in the identification of disease genes. Hence if a region of the mouse genome is mapped to high resolution, the information can be used to make predictions about the orthologous region of the human genome (and vice versa).

A comparative map of human chromosome 22 with mouse chromosomes shows conserved synteny to portions of mouse chromosomes 6, 16, 10, 5, 11, 8, and 15 (see Figure 1-5). The region of the mouse chromosome with conserved linkage to the CESCR has been defined as chromosome 6. T. Footz has assembled a mouse BAC contig spanning this region, and in collaboration with Dr. Bruce Roe at the University of Oklahoma , several of the BACs are currently being sequenced. At least nine genes show conserved linkage to mouse chromosome 6, including BID (Footz et al. 1998), ATP6E (Puech et al. 1997; Footz et al. 1998), and IL-17R (Yao et al. 1997).

- **Computer Analysis of Genomic Sequence**

As the HGP advances from its initial stages, more and more genomic sequence is becoming available through the sequencing efforts of many public and private sector centers. The most efficient approach to identify genes based on genomic sequence data is through the use of computer similarity searches of sequence databases and exon and gene prediction programs. The most useful database to identify putative coding sequences of a genomic sequence is the EST database (dbEST). An EST is generated by sequencing short segments (usually 200-300 bp) at the ends of cDNA clones. A considerable variety of EST projects, such as the IMAGE consortium, have generated massive numbers of EST sequences which are publicly available through the EST database. Sequencing both ends of a cDNA generates two ESTs: a 3' EST and a 5' EST. Because the 3'-untranslated sequence (3'UTS) of human genes usually falls within the 400 - 800 bp range, the 3' EST sequences are usually derived from the 3'-UTS. Not all EST cDNAs however, represent true genes. Some ESTs represent chimeric clones or pseudogenes, or some may be the result of mis-priming from non-coding genomic DNA, especially at a long stretch of poly (A) (when the primer used is oligo d(T)) which is often associated with some repetitive elements.

In order to identify matching ESTs, the genomic sequence can be searched for similarity to previously characterized human and non-human nucleotide sequences using BLASTN, or protein, using BLASTP. Not all genes will be represented by ESTs, however, because some genes will be in low abundance or not expressed at all in the cDNA libraries analyzed. This includes genes which show highly restricted tissue or cell expression patterns and genes which are expressed only at specific developmental stages. It is expected that the EST projects should be able

to identify sequences from approximately 80 % of all the human genes (Strachan & Read 1996).

A number of exon/gene prediction programs have been developed to predict coding sequences from genomic DNA. Gene modeling programs such as GENSCAN (Burge & Karlin 1997, 1998), FGENEH (Solovyev 1994), and Genie (Kulp et al. 1996) analyze large genomic sequences and identify genes based on the coding potential of the DNA by comparing it to the DNA and protein sequences in various sequence databases. Most exon prediction programs like GRAIL, MZEF (Zhang 1997), FEXH, and HEXON (Solovyev et al. 1994) are designed to scan a genomic DNA sequence to identify the locations of likely exons by screening for open reading frames (ORFs) and conserved sequences found at exon-intron junctions.

Once putative genes and exons are predicted, they must be confirmed using more direct methods. Scanning the EST database with the predicted coding region(s) will allow identification of any previously identified human or non-human genes or ESTs. If no ESTs (or full length cDNAs) are found, cDNA library screening, RT-PCR or RACE may be performed.

It is generally believed that in order to produce a near-complete transcription map, a number of gene identification techniques must be employed. Using any one of the techniques will usually identify only a subset of all the genes in a region. In this study a combination of exon trapping and genomic sequence analysis (at a later date) was applied.

22q Pericentromere: Rich in Repetitive DNA and Unprocessed Pseudogenes and Posing a Challenge to Transcript Mapping in the CESCR

The association between repetitive DNA and centromere structure is well known. Almost every species of animal studied to date harbours an array of both highly repetitive and moderately repetitive elements at the site of each chromosome's primary constriction. In humans, this repetitive DNA is usually composed of arrays of AT-rich tandomly repeated DNA sequences (alpha and/or beta satellite etc.) and interspersed repetitive DNA elements (SINEs , LINEs etc.). The presence of these repetitive DNA sequences in the centromeric regions of chromosomes are thought to favour the assembly of the kinetochore, thereby ensuring the timely and efficient meiotic and mitotic segregation of chromosomes (Haaf et al. 1992; Larin et al. 1994; Tyler-Smith et al. 1993). However, the presence of repetitive DNA, by itself, does not explain adequately the molecular basis of a functionally competent centromere. Barry et al. (1999) have presented the first complete sequence analysis of neocentromere DNA isolated from cytogenetic band 10q25.2. A neocentromere is a new centromere at a previously non-centromeric region on a chromosome arm (Barry et al. 1999). The formation of a neocentromere has been associated with numerous morphologically abnormal marker chromosomes which have lost their normal centromere through chromosomal rearrangements (Choo, 1997; Depinet et al. 1997). The regions of neocentromere formation have been proposed to be the sites of latent centromeres that can become activated through unknown epigenetic mechanisms (Choo, 1997, 1998; Brown & Tyler-Smith 1995; Karpen & Allshire 1997; Williams et al. 1998). These neocentromeres function as the site for kinetochore assembly and spindle fiber attachment in the same manner as a normal

centromere. Barry et al (1999) showed the absence of classic alpha-satellite repeat elements, beta-satellite, gamma-satellite, AT-rich sequences and a low frequency of interspersed repetitive DNA sequences within the neocentromere. These results suggest that highly repetitive DNA does not appear to be a prerequisite for centromere function. Csink and Henikoff (1998) have recently proposed a model which defines the functional relevance of centromeric repetitive DNA. They define a centromere as the latest replicating segment of DNA. They propose the idea that the accumulation of repetitive DNA at or near the centromere further retards replication, effectively fixing such sites in the genome to function more competently as a centromere. Therefore, the universality of repetitive DNA at centromeres is an evolutionarily derived state and not a precursor of centromere function.

If the centromere is a recruitment spot for repetitive DNA sequences, perhaps it may also function as a reservoir for the accumulation of duplicated genomic segments. Within the subtelomeric and pericentromeric portions of chromosomes, large tracts (several kb to hundreds of kb) of duplicated genomic segments that show a high degree of homology have recently been identified. These paralogous (sequence similarity due to duplication) fragments appear to carry complete or partial genomic structure of known genes, suggesting that they have been recently transposed (1-15 Million years ago (mya)) from other regions of the genome. Most of these segments were identified among sequences mapping to the pericentromeric regions of chromosomes 2p11, 10p11, 15q11, 16p11 and 22q11 (Eichler et al. 1996, 1997; Jackson et al. 1999; Hulsebos et al. 1996; Regnier et al. 1997; Wong et al. 1990; Potier et al. 1998; Ritchie et al. 1998). The term pericentromeric refers to a large transition zone that begins immediately distal to the alpha-satellite repeat and extends into the first

distinguishable cytogenetic G-band on either side of the centromere (Eichler et al. 1998).

Several paralogous genomic segments, or "duplicons" (Eichler et al. 1997), are known to reside within the pericentromeric region of 22q and therefore within the CESCR. These include fragments of the neurofibromatosis type 1 gene (NF1), the vonWillebrand factor gene (vWF), the γ-glutamyl transferase gene (GGT) family, a novel calcium-activated potassium channel regulatory subunit gene (KCNMB3) and the immunoglobulin kappa light-chain gene (Lotscher et al. 1986). All of the duplicons found in the 22q pericentromere represent truncated non-processed pseudogenes. Pseudogenes of this type have intact introns but either lack an essential 5' or 3' portion of the gene, and/or the exons have deleterious mutations which interrupt the normal translation of the gene to protein.

NF1 is a dominantly inherited genetic disorder involving the development of benign and malignant tumors in tissues derived from the neural crest. The NF1 gene is located at 17q11.2. Southern and FISH analysis has identified eleven NF1-related sequences on seven different chromosomes (2, 12, 14, 15, 18, 21, and 22) (Regnier et al. 1997). All of these NF1-related sequences represent unprocessed nonfunctional pseudogenes. The gene fragment on 22q11 contains exons 7-28, a region of approximately 100 kb (Regnier et al. 1997) and is located near D22S111 within the 2 Mb CESCR (See Figure 1-2) (Hulsebos et al. 1996). Comparison of the genomic sequence of exons 13-15 between NF1 and the 22q11 pseudogene show approximately 8 % divergence. This sequence divergence corresponds with an ancestral gene duplication event about 2-33 mya.

Similar duplication and transposition events may account for the paralogous segments of the ALD gene within the pericentromeric regions of chromosomes 2p, 10p, 16p, and 22q (Eichler et al. 1997). Like the NF1-related sequences, these duplicons are truncated unprocessed pseudogenes composed of a 9.7 kb genomic fragment encompassing exons 7-10 and originating from the active ALD gene located at Xq28 (Eichler et al. 1997). There has been approximately 4 % sequence divergence between the active copy and the 22q duplicon. Mutations in the ALD gene at Xq28 cause a relatively common X-linked neurodegenerative disorder.

The CESCR also harbours a vWF unprocessed and nonfunctional gene fragment. The active vWF gene is composed of 52 exons encompassing approximately 189 kb and is located at 12p13.3. Mutations in this gene cause a form of hemophilia known as vonWillebrand disease (Eikenboom et al. 1994). The pseudogene on 22q spans a region of approximately 30 kb near D22S50 (See Figure 1-2) and contains exons 23-34 (Mancuso et al. 1991). The sequence of the 22q gene fragment has diverged about 3 % from the active vWF gene.

Riazi et al. (submitted) have identified a truncated unprocessed pseudogene within the CESCR, KCNMB3L, located within the vicinity of D22S795. The active copy of this gene is located at 3q26.3-27 and contains three exons, two of which (exons two and three) were duplicated and transposed to 22q11. The sequence divergence between the active gene and the 22q pseudogene is 2.6 % (Riazi Ph.D.), however, the sequence of the 22q pseudogene has maintained an ORF.

Another gene which has been duplicated and transposed to the 22 pericentromeric region is the γ-glutamyl transferase gene family. There may be up to seven copies of GGT within 22q11. Some of these may

encode for transcripts (Courtay et al. 1994) however the copy in the CESCR, GGT12, is a 5' truncated gene fragment, located in the vicinity of D22S795 (See Figure 1-2). Other low copy number repeats besides GGT, have been found within 22q11, including D22S131, D22S207, and D22S287E (McTaggart 1997). It is possible that the deletion and duplication syndromes, DGS/VCFS and CES, are due to the instability of the 22q11 region. This instability may be due to the misalignment and unequal crossing over between these repeats, similar to that on 17p causing Charcot-Marie-tooth disease type 1A (CMT1A) and heredity neuropathy with liability to pressure palsies (HNPP) (Pentao et al. 1992; Strachan & Read 1996).

The presence of these paralogous genomic segments within the CESCR poses a challenge to transcript mapping in this region. Exons or cDNAs corresponding to these gene fragments may be easily identified by conventional gene identification techniques such as exon trapping, cDNA selection as well as computer-aided gene modeling programs. Much time and effort can be spent in the characterization of such exons/cDNAs before it is evident that they correspond to a duplicated gene fragment. It is likely that the number of duplicons within the 22q pericentromere correlates with proximity to the centromere (Eichler et al. 1996, 1997, 1998). Therefore we expect that the 1 Mb proximal CESCR will harbour many of these duplicons and be extremely gene-poor (Minoshima et al. 1998), whereas the distal CESCR may contain fewer duplicons and be more gene-rich (McDermid, personal communication).

The presence of multiple duplicated, unprocessed pseudogenes in the genome is not without clinical significance. Many of the genes duplicated, such as ALD and vWF, are associated with diseases. Related gene fragments could produce difficulty in their molecular diagnoses and therefore the study of these paralogous segments is of great significance.

The "pericentromeric plasticity" seen with chromosome 22 and others may be of evolutionary significance. The duplication and transposition of various gene fragments to pericentromeric regions may create a reservoir of genes in the genome with potentially new functions. Exons from various gene fragments, in the presence of promoter-type sequences, may be transcribed into a novel gene product. Such events, if they did occur, could accelerate an organism's adaptability and allow for rapid genetic differences within a species (Eichler et al. 1997). Such a "gene" would not have an orthologue in any model organism (Yeast, mouse, *Drosophila*, *C.elegans*) since it would have evolved after the divergence of man. Such a human-specific gene would be difficult to study without the availability of model organisms. It is also tempting to speculate that such a fusion gene product(s) may be the cause of deletion/duplication syndromes located near some centromeres.

RESEARCH OBJECTIVES

The main purpose of this project was to isolate and characterize genes located within the 1 Mb distal CESCR. A traditional gene identification technique, exon amplification, was applied to a number of PAC/BAC/cosmid clones within the CESCR by Dr. Ali Riazi. A subset of the isolated putative exons were characterized in this study.

The exons were characterized by sequencing, Southern hybridizations and computer database searches. To identify putative genes within the CESCR ESTs or cDNAs containing the isolated exons were identified by screening cDNA libraries and by performing BLASTN and TBLASTX searches of the EST database. The isolated cDNAs were characterized by sequencing and Southern and Northern hybridizations. To try to complete the cloning of these putative genes, 5' and 3' RACE and RT-PCR were performed. These results are presented in Chapter 3.

To further characterize the putative genes discovered by exon trapping, genomic sequence (available at a later date) was analyzed for the presence of computer predicted genes and exons. The predicted elements were compared with the exons (and associated cDNAs) obtained through exon trapping. The computer predicted gene elements were used to further clone the identified genes.

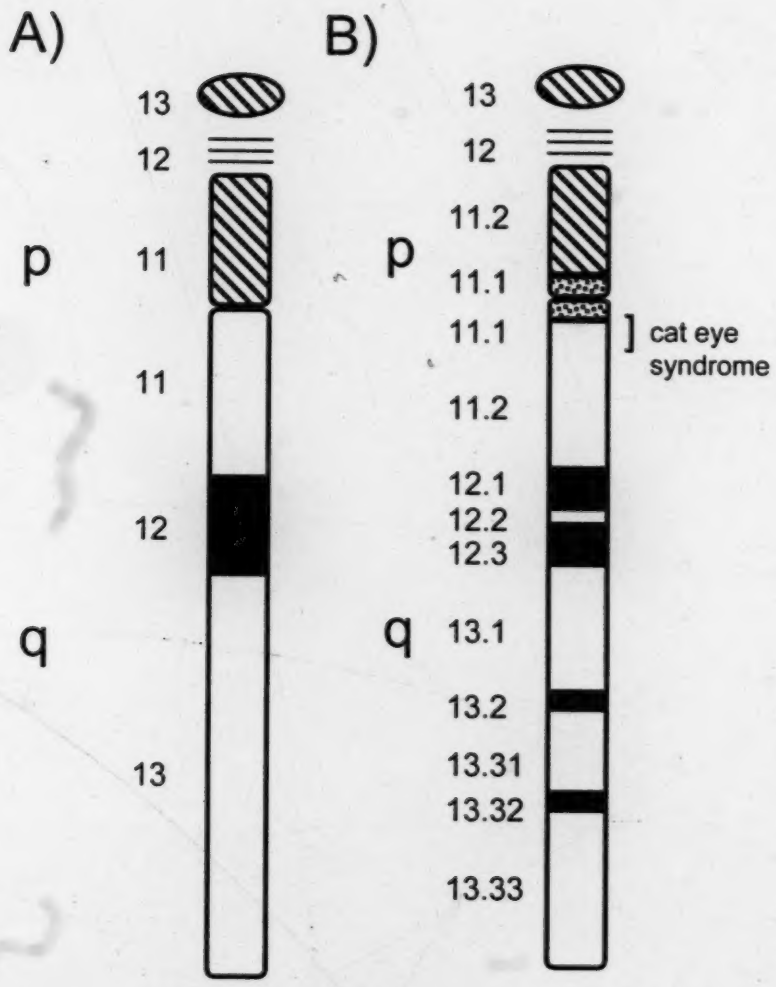


Figure 1-1: Ideograms of chromosome 22

at the **A)** 400 band level and **B)** 850 band level (modified from Harden and Klinger ISCN 1985). Duplication of the most proximal region of 22q causes cat eye syndrome (CES).

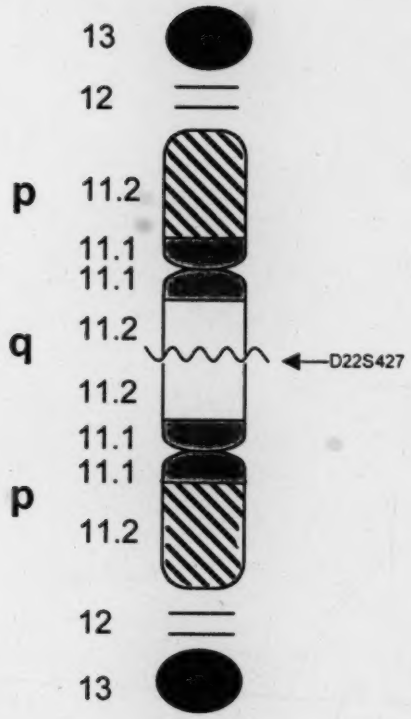
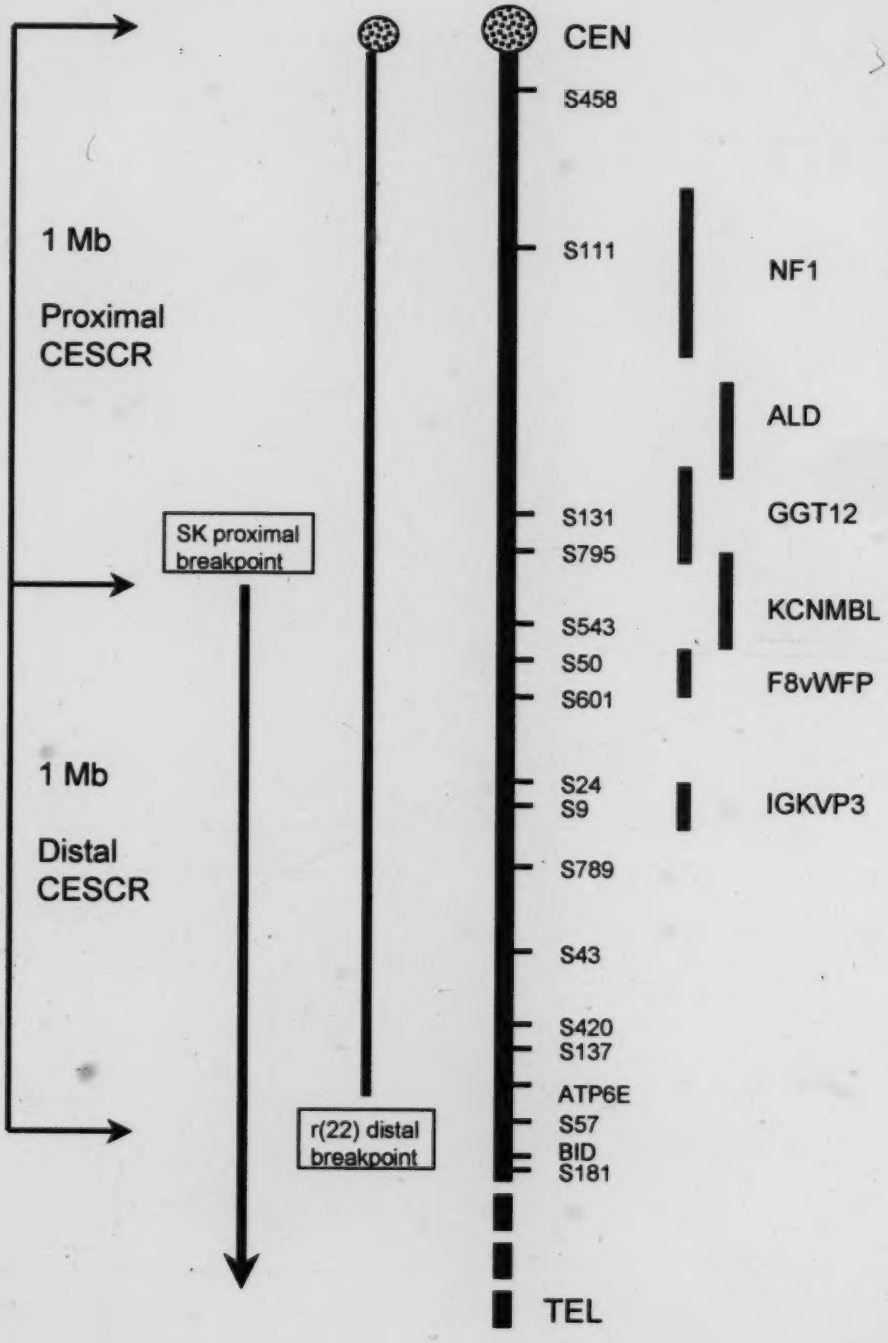


Figure 1-2: A Cat Eye Chromosome
Individuals with a supernumerary cat eye chromosome will have four copies of the region 22pter-22q11.2. The chromosome shown is a type I, with breakpoints near the locus D22S427.

Figure 1-3 : The Cat Eye Syndrome Critical Region (CESCR).

The smallest region of proximal 22q11.2 required to be duplicated to result in the CES phenotype is a 2 Mb region extending from the centromere to locus D22S57. This distal boundary was delineated by a patient with a small supernumerary r(22) chromosome. The locus D22S57 was not present on the r(22) but the next most proximal probe, ATP6E, was present in four copies in this individual. This patient presented with all of the cardinal features of CES. The proximal breakpoint of an individual carrying an interstitial duplication was found to be between loci D22S795 and D22S543, a distance of approximately 1 Mb from the centromere. This patient presented with an incomplete phenotype, lacking iris coloboma and anal atresia. Thus the CESCR was divided into proximal and distal halves. Comparing the duplication overlap between these two patients, the region of focus has been narrowed to the 1 Mb distal CESCR. This map is drawn roughly to scale based on McDermid et al. 1996 and Collins et al. 1995.

The estimated location of six reported pseudogenes are represented by red bars to the right of the map. Their tentative locations are based on McDermid et al. 1996, Collins et al. 1995, 1997, Hulsebos et al. 1996 and McDermid , personal communication.


1 Mb
Proximal
CESCR
1 Mb
Distal
CESCR
SK proximal
breakpoint
r(22) distal
breakpoint
CEN
S458
S111
S131
S795
S543
S50
S601
S24
S9
S789
S43
S420
S137
ATP6E
S57
BID
S181
TEL
NF1
ALD
GGT12
KCNMBL
F8vWFP
IGKVP3

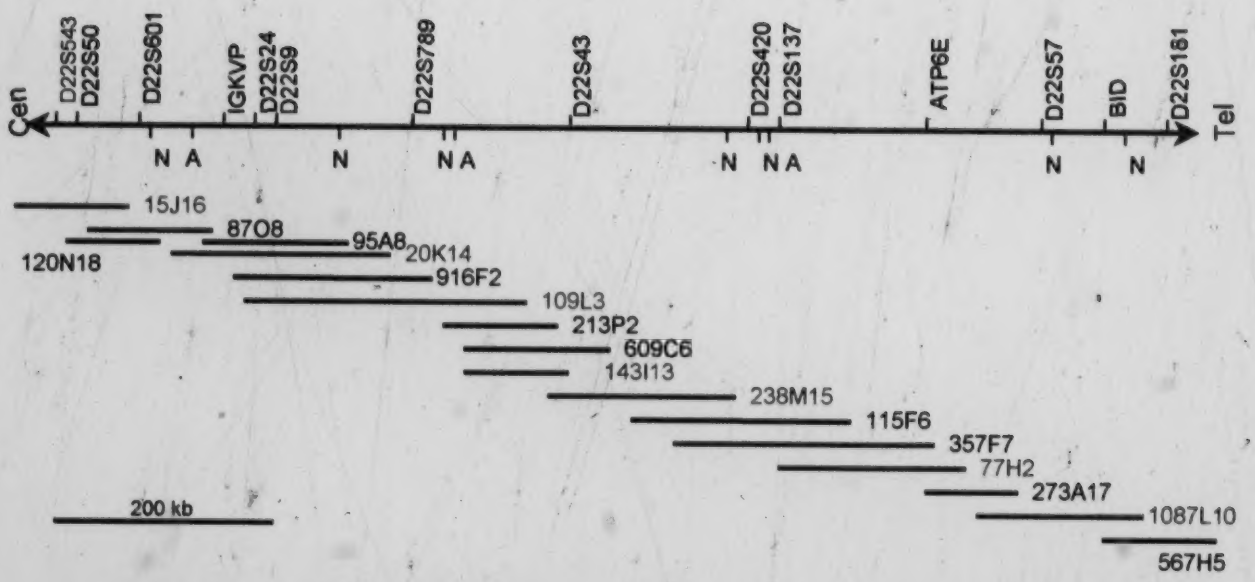


Figure 1- 4: A partial physical map of the CESCR (redrawn from Johnson et al. 1999)
STSs (or known genes) are indicated above the thick horizontal double-arrowed line. A pulsed field gel electrophoresis map is shown below (N = *Not* I A = *Asc* I). Clones that have been or will be sequenced are represented by green lines. Clones which have complete or near complete sequence available as of August 31,1999 have their address written in red. The map is drawn approximately to scale.

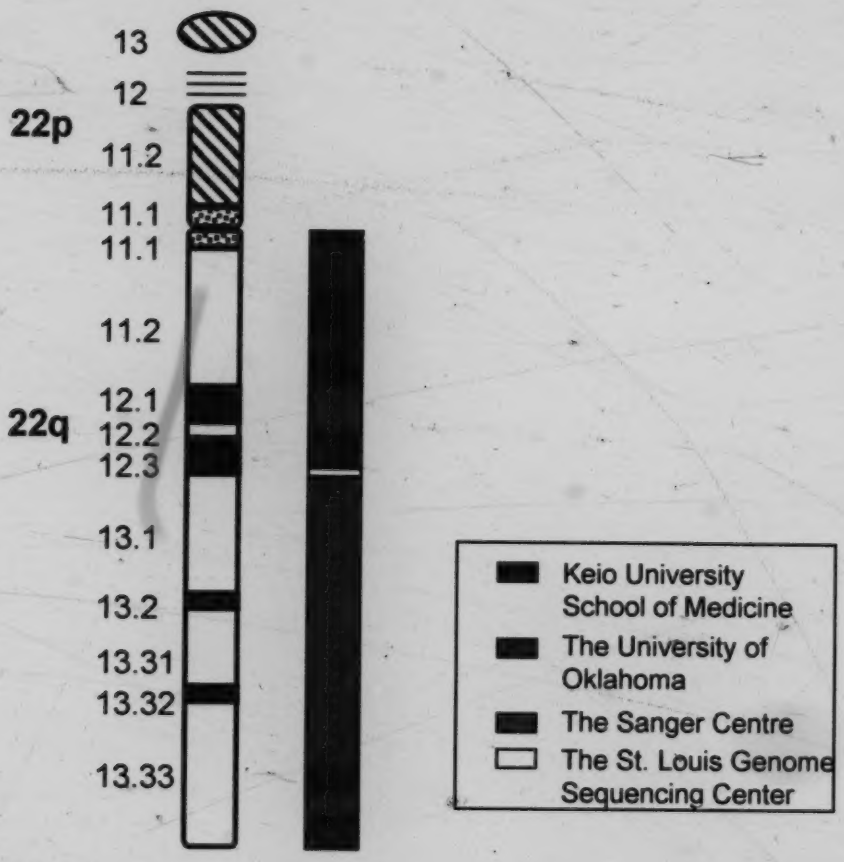


Figure 1- 5: The Chromosome 22 Sequencing Consortium

The colored vertical bar adjacent to the chromosome 22 ideogram represents the major sequencing centers responsible for regional sequencing. The color key representing each sequncing center is located at the right. Genomic clones within the CESCR are being sequenced at the University of Oklahoma.

(modified from http://www.sanger.ac.uk/HGP/Chr22/Mouse/).

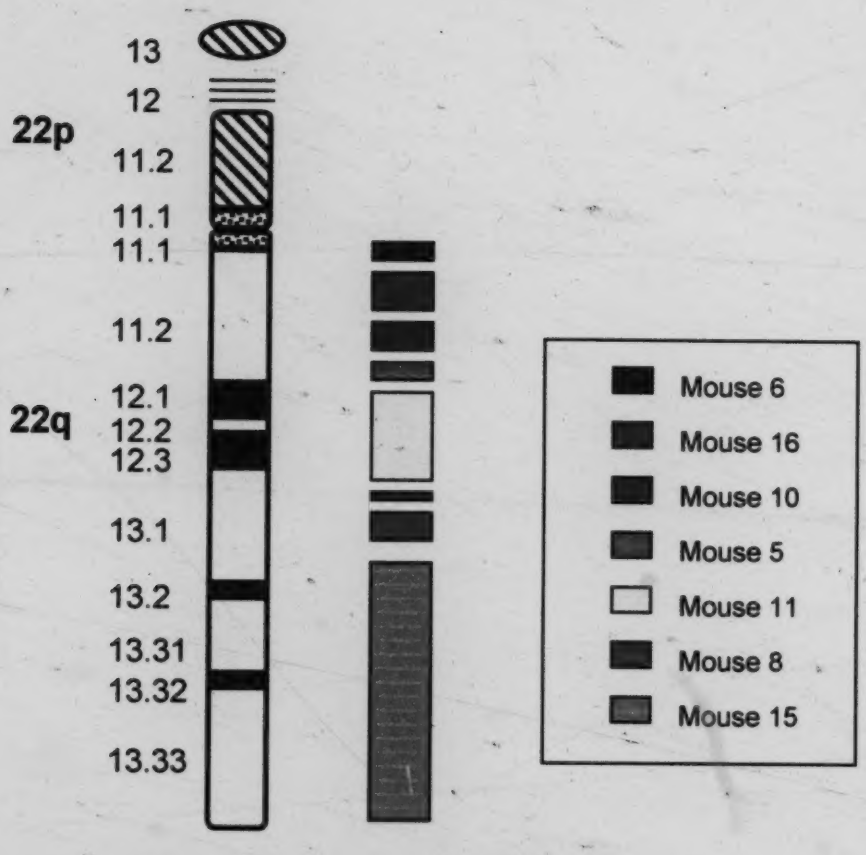


Figure 1-6: Human - Mouse Comparative Map

The colored vertical bar adjacent to the chromosome 22 ideogram represents the syntenic mouse regions. Chromosome 22 is syntenic to portions of mouse chromosomes 6, 16, 10, 5, 11, 8, and 15. These regions are color coded as defined to the right. The region of the mouse chromosome with conserved linkage to the CESCR has been defined as chromosome 6. (modified from http://www.sanger.ac.uk/HGP/Chr22/Mouse/)

PHENOTYPE	% FREQUENCY
Preauricular Malformations	80 %
Anal Atresia (imperforate anus)	75 %
Coloboma	50 %
Downslanting Eyes	50 %
Heart Defects (TAPVR, TOF)	45 %
Renal Defects (absent or hypoplastic kidney)	45 %
Hypertelorism	35 %
Low Set / Malformed Ears	35 %
Mental retardation (mild to moderate)	30 %
Skeletal Defects	15 %
Genital Defects	10 %

Table 1-1: Frequency of CES Congenital Defects

The most common forms of defects are indicated in brackets. TAPVR refers to total anomalous pulmonary venous return and TOF refers to Tetrology of Fallot. (Modified from Mears 1995).

CHAPTER 2: MATERIALS AND METHODS

Exon Trapping

Exon trapping was performed by Dr. Ali Riazi on two BACs (95A8, 233A2), three PACs (109L3, 238M15, 120N18) and 27 cosmid clones using the GIBCO BRL Exon Trapping System. Exon trapping was performed according to the instructions included with the kit with some modifications (Riazi, Ph.D. thesis, 1998). Figure 2-1 illustrates a brief overview of this technique.

Preparation of Bacterial or P1-derived Artificial Chromosome (BAC or PAC) DNA

Large-scale DNA preparations were performed according to the QIAGEN® plasmid purification procedure. Cultures were grown in 5 ml of Luria-Bertani medium supplemented with 12.5 µg/ml chloramphenicol for 8 hours at 37°C in a shaking incubator. This culture was used to inoculate a 1 litre culture of chloramphenicol-supplemented LB which was subsequently incubated shaking at 37°C overnight. The bacterial cells were harvested by centrifugation at 6000 rpm in a Sorval GSA rotor for 15 minutes at 4°C. All visible traces of the supernatant were removed by inverting the open centrifuge tube until all medium had drained. The bacterial pellet was resuspended completely in 10 ml of Buffer P1 (50 mM Tris-HCl, pH 8.0; 10 mM EDTA; 100µg/ml RNase A). 10 ml of Buffer P2 (200 mM NaOH; 1% SDS) was then added, the solution was mixed gently by inverting the tube several times, and incubated at room temperature for 5 minutes. To precipitate proteins, chromosomal DNA, and cellular debris, 10 ml of Buffer P3 (3 M potassium acetate, pH 5.5) was added.

The solution was mixed immediately by inversion and incubated on ice for 20 minutes. The precipitated debris was pelleted by centrifugation at 11000 rpm for 30 minutes at 4°C. The cleared supernatant was removed to a clean tube. A QIAGEN-tip 500 was equilibrated by applying 10 ml of Buffer QBT (750 mM NaCl; 50 mM MOPS, pH 7.0; 15% ethanol; 0.15% Triton X-100) which emptied through the column by gravity flow. The cleared supernatant was filtered over a prewetted, folded filter (Whatmann paper) and allowed to enter the tip-500 by gravity flow. To remove any remaining contaminants, such as traces of RNA and protein, the tip-500 was washed twice with 30 ml of Buffer QC (1 M NaCl; 50 mM MOPS, pH 7.0; 15% ethanol). The DNA was eluted from the column with 15 ml of 50°C Buffer QF (1.25 M NaCl; 50 mM Tris-HCl, pH 9.0; 15% ethanol) into a clean tube. Precipitation of the DNA was performed using 0.7 volumes of room-temperature isopropanol. The sample was centrifuged at 11000 rpm for 30 minutes at 4°C. After centrifugation, the pellet was washed with 5ml of 70% ethanol to remove residual salt and replace the isopropanol with the more volatile ethanol. The sample was re-centrifuged at 10000 rpm for 5 minutes at 4°C. After careful removal of the ethanol, the pellet was air-dried for approximately 15 minutes and then dissolved in an appropriate volume of TE buffer (10 mM Tris-HCl, pH 8.0; 1 mM EDTA). The concentration of the resultant DNA was assessed by analysis on an agarose gel.

Preparation of Plasmid DNA

For small-scale preparations of plasmid DNA, the method of choice was the alkaline lysis procedure as described in Sambrook et al. (1989). A single colony was used to inoculate 5 ml of LB supplemented

with 50 µg/ml Ampicillin. The culture was grown overnight at 37°C in a shaking incubator. This method typically yielded 20-25 µg of plasmid DNA.

Southern Blotting and Hybridization of DNA probes

Typically 1 µg of plasmid DNA or 3 µg of BAC/PAC or genomic DNA was digested with the desired restriction enzyme as per manufacturer's instructions. Electrophoresis of the digested DNAs was performed on a 0.7-1.0 % agarose (GIBCO BRL) gel in 1X TBE (0.45 M Tris-HCl; 0.45 M Boric Acid; 2 mM EDTA; pH 8.4) at various voltages depending on the size of the gel. Plasmid and BAC / PAC DNA was transferred to GeneScreen *Plus* ® (Du Pont) by the alkaline transfer protocol suggested by the manufacturer. Genomic DNA was transferred to GeneScreen *Plus* ® (Du Pont) by the method of Southern (1975). The gel was agitated in denaturing solution (0.5 M NaOH; 1.5 M NaCl) for 30 minutes, followed by two, 30 minute agitations in neutralization solution (0.5 M Tris pH 7.5; 3M NaCl). The capillary transfer was performed in 10X SSC. After the transfer was complete, the membrane was soaked in 0.4 M NaOH for 5 minutes, to denature the DNA. The membrane was then neutralized by shaking in 0.2 M Tris pH7.5; 2X SSC for 5 minutes.

To make DNA probes for hybridization to Southern blots, the DNA fragment was isolated from low-gelling temperature agarose (SEAPLAQUE® FMC Inc.) and labeled by the random priming method (Feinberg & Vogelstein, 1984). Approximately 50-100 ng of the purified DNA fragment was boiled for 10 minutes to denature the DNA and then added to a mixture of 10 µl of oligo labeling buffer (1 M Tris, pH7.5; 1 M $MgCl_2$; 100

mM dATP, dGTP, dTTP; 2 M Hepes, pH 6.6; 90 A_{260} units/ml oligonucleotides; 55mM β-mercaptoethanol), 2 µl 10X BSA (BRL or NEB) and 11 µl sterile water. To this cocktail was added 5 µl (α-^{32}P)dCTP (Amersham) and 2µl (6435 units/ml) of purified DNA Polymerase I Klenow Fragment (Pharmacia Biotech). Following incubation for 2 hours at 37° C, the mixture was passed through a column containing Sephadex G-50 (Pharmacia Biotech) in order to remove unincorporated nucleotides. The resulting probe was added to 250 µl sonicated human placental DNA (2.5 mg/ml) and 42 µl $NaPO_4$ (1M; pH 7.0) and the mixture was boiled for 10 minutes, and then incubated at 65° C for 30-60 minutes in order to preanneal any repetitive DNA elements (Litt & White, 1985).

Southern blots were prehybridized for at least 30 minutes in 10 ml of hybridization solution {(10 % SDS; 1 M EDTA; 1 M $NaPO_4$; 5X Denhardt's), slight modification of Church and Gilbert, 1984} in a rolling bottle hybridization oven (Tyler). The probe was then added to the blot and hybridized for 16-24 hours. The following day the blot(s) was/were washed as follows: 2 X 10 minutes at room temperature (1.5 X SSC; 0.2 % SDS); 2 X 10 minutes at 65° (0.2 X SSC; 0.2 % SDS); and if necessary, in 10 minute intervals at 65° (0.1 X SSC; 0.2 % SDS). The blot was sealed in a plastic bag and exposed to Kodak X-Omat film at -70°C for the desired amount of time (usually a few hours to several days).

When a blot was subjected to hybridization with a different probe, the previous probe was stripped off the blot by boiling the membrane for approximately 30 minutes in a solution of 1 % SDS; 0.1 X SSC.

Northern Blots and Hybridization

Human multiple tissue Northern blots were purchased from either CLONTECH Laboratories, Inc. or Invitrogen® . Each Northern blot contains approximately 2 µg of poly A^+ RNA per lane from various different human adult or fetal tissues.

To make DNA probes for hybridization to Northern blots, the DNA fragment was isolated from low-gelling temperature agarose and labeled using the Strip-EZ ™ DNA kit (Ambion). Approximately 50-100 ng of the purified DNA fragment was boiled for 10 minutes to denature the DNA and then added to a mixture of 2.5 µl 10X Decamer Solution, 2.5 µl 10X dCTP, 5 µl 5X buffer. To this mixture was added 5 µl (α-^{32}P)dATP (Amersham) and 1µl of Exonuclease-free Klenow. Following incubation for 10 minutes at 37° C, the mixture was passed through a column containing Sephadex G-50 in order to remove unincorporated nucleotides. The resulting probe was preannealed as described previously.

Northern blots were prehybridized for at least 30 minutes in Northern Hybridization solution (50 % Formamide; 5 X SSPE; 10 X Denhardt's; .07 % SDS; 0.4 mg / ml Herring sperm DNA) in a rolling bottle hybridization oven. The probe was then added to the blot and hybridized for 16-24 hours. The following day the blot was washed as follows: 2 x 15 minutes at room temperature (2 X SSC; 0.1 % SDS); 2 x 10 minutes at 42-50 °C (0.1 X SSC; 0.1 % SDS), checking for signal with a Geiger counter between washes. The blot was sealed in a plastic bag and exposed to Kodak X-Omat film at -70 °C for the desired amount of time (usually 1-7 days).

When a blot was subjected to hybridization with a different probe, the previous probe was stripped off the blot using Ambion's instructions. The blot was washed for 10 minutes at 68° in 10 ml of a 1 X mixture of Probe Degradation Dilution Buffer and DNA Probe Degradation Buffer (all stripping reagents were provided with the kit). The first wash was poured off and 10 ml of a mixture of 1 X Blot Reconstitution Buffer and 0.1 % SDS was added. The blot was washed for a further 10 minutes at 68°. This standard probe removal protocol removed "greater than 95 % of the hybridized strippable probe" (Ambion).

Cycle Sequencing

Exons and cDNAs were sequenced using a Thermo Sequenase radiolabelled terminator cycle sequencing kit (Amersham Life Science) according to the manufacturer's instructions. Approximately 0.25 µg of the DNA to be sequenced was added to 2 µl reaction buffer (260 mM Tris-HCl, pH 9.5; 65 mM $MgCl_2$), 1 µl primer (1 pmol/µl), 2 µl Thermo Sequenase polymerase (4 U/µl) and sterile water (to 20 µl) and subsequently aliquoted into 4 X 4.5 µl reaction mixtures. Termination mixes were prepared by combining 2 µl dGTP termination master mix (7.5 µM dATP, dCTP, dTTP, dGTP) and 0.5 µl of [α-^{33}P] ddNTP (G, A, T, C - one of each per sequence; 450 µCi/ml). Each termination mix was added to it's corresponding reaction mix (G with G and so on). The tubes were capped and placed in a Progene thermal cycler (Techne) using the following parameters: Denaturation 95 °C, 30 sec; Annealing X °C, 30 sec (depending on the length and GC content of the primer); Extension 72 °C,

60 sec; for 10 cycles. When cycling was finished, 4 µl of stop solution (95 % formamide; 20mM EDTA; 0.05 % bromophenol blue; 0.05 % xylene cyanol FF) was added to each reaction tube. The reaction was heated to 75 °C for 5 minutes and 6 µl was loaded immediately onto an 8 % denaturing polyacrylamide gel { 5.32 g Acrylamide; 0.28 g bis-acrylamide; 32 g urea; 3.5 ml 20X glycerol tolerant buffer; sterile water to 70 ml; filter and de-gas. When ready to pour add 500 µl 10 % ammonium persulfate and 12.5 µl TEMED (N,N,N',N'-tetramethylethylenediamine)}. The gel was run at 55 Watts for 1.5-7 hours and dried under vacuum at 80 °C for 2 hours. Autoradiography was performed using Kodak BioMax MR film at -70 °C for the desired amount of time (usually 1-3 days).

Screening Human cDNA Libraries

Various cDNA libraries were screened to identify cDNA clones containing previously identified exons; a list of the libraries used is presented in Table 2-1.

For plating of these libraries, the bacterial host cell used was *E. coli* XL1-Blue (recA1, endA1, gyrA96, thi-1, hsdR17, supE44, rela1, lac,[F'proAB, laclqZDM15, Tn10 (tet^r)]. XL1-Blue were grown in LB containing 0.2 % maltose and 10 mM $MgSO_4$. After growing overnight at 37 °C, the cells were pelleted and resuspended in 10mM $MgSO_4$ to an OD_{600} of 0.5. The libraries were titered on LB plates using various dilutions of the original stock and by conventional methods (Sambrook et al., 1989). Aproximately 30,000 phage were plated on each 150mm x15mm plate, and 20-25 plates were made. The plates were incubated for aproximately 14 hours at 37 °C and then cooled at 4 °C for several hours.

Plaque lifts were performed using Hybond-N (Amersham LIFE SCIENCE) nylon membranes as per manufacturer's instructions. The membrane was placed on the plate for one minute and marked with pinholes for future orientation. The membrane was then placed DNA-side up, onto a Whatmann-soaked pad of denaturation solution (1.5 M NaCl; 0.5 M NaOH) for three minutes. This was followed by 2 X 3 minute neutralization steps (1.5 M NaCl; 2 M Tris, pH 7.5). Finally the membrane was agitated in 2 X SSC for 5 - 10 minutes (to remove any cellular debris adhering to the membrane). The DNA was fixed to the membranes by baking for 2 hours at 80 °C in a vacuum oven.

Hybridization of radioactively-labeled DNA probes was performed in the manufacturer-recommended solution (5 X SSC; 5 X Denhardt's solution; 0.5 % SDS) for 16-20 hours at 65 °C. The region around the primary positive plaque was isolated from the original plate, re-plated at a low density on LB plates, plaque-lifted and re-probed. This procedure was repeated until a single positive plaque could be isolated. *In vivo* excision of the pBluescript SK$^-$ plasmid from the λZAPII vector was performed using R408 helper phage (Stratagene) as per manufacturer's instructions.

Polymerase Chain Reaction

Polymerase chain reactions were performed using a PTC-100 ™ programmable thermal controller (MJ Research Inc.). Usually 40 µl reactions were set up in 1 X PCR buffer (500 mM KCl; 100 mM Tris, pH 7.5; 15 mM $MgCl_2$), 10 pmol of each primer (GIBCO BRL), 10mM dNTPs, the desired amount of template DNA, and sterile water to 40 µl. The

reactions were heated to 95 °C for 2 minutes and held at 80 °C while 0.5 µl of *Taq* polymerase (Department of Microbiology, University of Alberta) was added to each (this technique was referred to as "hotstart" PCR). The PCR conditions typically consisted of an initial denaturation at 95 °C for 1-2 minutes followed by 25-35 cycles of annealing (T_m- 3 to 5 °C for 30 seconds), extension (72 °C for 30 seconds / 0.5 kb) and denaturation (95 °C for 30 sec) followed by a final annealing and 10 minute extension at 72 °C.

When a larger PCR product (≥ 1.5kb) was desired, the *TaqPlus Precision*™ PCR System (Stratagene®) was used as per manufacturer's instructions. Typically 50 µl reactions were set up in 1 X PCR buffer (provided with kit), 10 pmol of each primer, 25 mM dNTPs, the desired amount of template DNA, and sterile water to 50 µl. Hotstart PCR was performed and 0.5 µl (5 U/µl) of *TaqPlus Precision* polymerase was added to each. The PCR conditions were as previously stated. To test for DNA contamination of any of the components of the PCR reaction, a water control consisting of all the components except the template DNA was also performed. Table 2-2 lists all the primers used in the described experiments.

Reverse Transcription PCR

Reverse transcription PCR (RT-PCR) was performed on 2-5 µg of total RNA isolated from various human tissues (by Dr. Andrew Wong or Dr. Valerie Trichet) including adult heart, liver, kidney, skeletal muscle and others, as well as fetal brain and two cell lines, CaCo (a colon carcinoma

cell line) and HeLa (a cervical cancer cell line). The RNA was added together with 20 pmol of a gene-specific primer and DEPC-treated water to 12 µl followed by incubation at 70 °C for 5 minutes to denature the RNA. Following 1 minute on ice, 4 µl of first strand buffer (250 mM Tris, pH 8.3; 375 mM KCl; 15 mM $MgCl_2$) , 2 µl 0.1 M DTT and 1 µl 10 mM dNTPs was added. Following incubation at 42 ° C for 2 minutes, 1 µl (200 Units) of SUPERSCRIPT II Reverse Transcriptase (GIBCO BRL) was added. The reaction was incubated at 42 °C for 30 minutes followed by 5 minutes at 55 ° C. To degrade the remaining mRNA, 1 µl (3.8 U/µl) of RNase H (GIBCOBRL) was added and the reaction incubated at 55 ° C for 10 minutes. Typically 2-5 µl of the reverse transcription reaction was used in a PCR reaction. Negative control reverse transcription was performed as above without the addition of SUPERSCRIPT II Reverse Transcriptase.

Subcloning

Subcloning was performed in pBluescript II SK(+/-) when it was necessary for simplified sequencing or if a particular restriction fragment was needed in a plasmid vector. The restriction fragment of interest was isolated from a 1X TAE-agarose gel and purified using the GeneClean system (BIO 101 Inc.). The purified fragment was ligated to the desired amount of appropriately digested pBluescript in 5 X T4 DNA ligase buffer, 1 µl T4 DNA ligase (1 U/µl; GIBCOBRL) and sterile water to 10 µl. The reaction was incubated at room temperature for 1-3 hours and then half of the reaction was transformed into competent XL1-Blue cells and plated

on LB plates supplemented with Ampicillin (50 μg/μl), X-gal (800 μg/μl) and IPTG (800 μg/μl).

When it was desired, PCR products were subcloned using the pGEM ®-T Easy Vector Systems (Promega) as per manufacturer's protocols. Transformants were selected for as described above.

Rapid Amplification of cDNA ends (RACE; 3' and 5')

Rapid amplification of cDNA ends (Frohman et al., 1988) was performed using Marathon-Ready™ cDNA kits (CLONTECH Laboratories, Inc.) which are essentially premade, uncloned libraries of adaptor-ligated double-stranded cDNA ready for use as a template in either 5' or 3' RACE reactions (see Figure 2-2 for a brief overview of this technique). The kits used were prepared from poly A^+ RNA from either human adult heart or brain. A primary RACE was performed using an internal gene-specific primer and the Marathon Adaptor Primer (AP1), followed by a nested PCR using a nested gene-specific primer and the nested Marathon Primer (AP2). All RACE reactions were performed as previously described using the *TaqPlus Precision*™ PCR System (Stratagene®). Typically 1 μl of the primary PCR was used as the template for the nested PCR. A portion of the RACE reaction was examined on an agarose gel and transferred to a nylon membrane as previously described. This Southern blot was hybridized with a cDNA probe containing the most 5' known portion of the gene. Once the band(s) of interest were determined, they were gel-purified and cloned into the pGEM ®-T Easy vector and sequenced.

Database Searching and Genomic Sequence Analysis

A large number of publically available programs are particularly useful for identifying similarities between experimental sequences (e.g. cDNA, exon (DNA and/or protein format)) and large databases of publicly-available, partially-sequenced cDNAs (expressed-sequence-tags or ESTs), genomic, or known gene sequences. Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) is a search algorithm employed by various programs designed for sequence similarity searching. These programs ascribe significance to their findings using the statistical methods developed by Karlin and Altschul (1990, 1993).

Typically, exon, cDNA, or genomic sequences were submitted to the BLASTN/nr program, which searches for similarities to all DNA sequences in the database such as sequence-tagged-sites (STSs), CpG islands, known genes or genomic sequence. Subsequently the experimental sequence was submitted to BLASTN/dbest, which identifies similarities to human and non-human ESTs. BLASTP, BLASTX, TBLASTX or BEAUTY (BLAST Enhanced Alignment Utility; Worley et al., 1995) searches were performed to identify protein homology and the presence of any known functional domains. The BLAST series of programs used were accessed either through the Baylor College of Medicine's Search Launcher (http://kiwi.imgen.bcm.tmc.edu:8088/search-launcher/launcher.html) or through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). All searches were performed using the default parameters of the search program (see BLAST Help Manual at http://www.ncbi.nlm.nih.gov/BLAST/blast_help.html). Any ESTs matching the query sequence with close to 100 % identity were purchased from Genome Systems Inc. and further characterized.

To analyze partial or complete genomic sequence (obtained from Dr. Bruce Roe at http://www.genome.ou.edu/hum_totals.html) a series of programs available over the Internet were employed (see Figure 2-3 for an illustration). The genomic sequence was occasionally submitted to the Genome Annotation and Information Analysis server (GAIA, http://daphne.humgen.upenn.edu:1024/gaia/index.html). GAIA is a data analysis and storage system for genomic sequence and its annotation. GAIA accepts raw genomic sequence and automatically adds annotation such as GRAIL-predicted exons, EST and STS hits and repetitive elements. GAIA was also used to format partial genomic sequence into numbered, columnar lines.

The genomic sequence was also submitted to various gene and exon prediction programs. The most reliable and frequently used gene prediction program was GENSCAN (http://bioweb.pasteur.fr/seqanal/interfaces/genscan.html or http://ccr-081.mit.edu/GENSCAN.html; Burge & Karlin, 1997). GENSCAN is useful for the identification of complete gene structure in genomic DNA including exons, introns, promoter and poly-adenylation signals. GENSCAN outperforms some other gene prediction programs in that it recognizes partial genes as well as multiple genes in a single sequence on either or both DNA strands. GENSCAN is based on a probabilistic model of genomic sequence composition (gene structure) and therefore is able to assign a probability that a particular exon, for example, is correct. This probability is defined as the sum of the probabilities under the model of all possible gene structure descriptions which contain the exact exon in the correct reading frame. These probabilities provide a useful quantitative guide to the likelihood that a given exon is correct. For example, a study done by Burset & Guigó (1996) on a set of 570 vertebrate gene sequences found that a probability between 0.75 and 0.9

indicates a 74.8 % chance that the predicted exon is exactly correct (i.e.: the predicted exon exactly matched the true exon precisely; both endpoints were correct) and an 8 % chance that the predicted exon was wrong (i.e.: the predicted exon does not overlap the true exon at all). Some implications from the Burset & Guigó (1996) study are: a P>0.99 indicates that the exon will almost always be correct, a 0.99>P>0.5 indicates that the exon will be correct most of the time and if P<0.5 then this prediction is not reliable. GENSCAN more accurately predicts internal exons than initial or terminal exons and it predicts exons more accurately than poly-adenylation or promoter signals (GENSCAN Web Server http://gnomic.stanford.edu/GENSCANW.html, and Personal experience). One drawback of the GENSCAN program is that it can predict exons within repetitive elements (especially L1 elements), so it may be wise to submit only repeat masked sequences. Other gene prediction programs occasionally used include FGENEH and Genie (available via the BCM search launcher).

The genomic sequence was submitted to exon prediction programs including GRAIL, MZEF, FEXH, and HEXON (all available via the BCM search launcher). Of these programs, MZEF and GRAIL were found to be the most useful and frequently used. GRAIL2 (Gene Recognition and Assembly Internet Link) defines an exon as having an open reading frame bounded by the correct splicing acceptor and donor sites. GRAIL2 finds approximately 91 % of all coding regions but is less efficient at predicting exons less than 100 bp in size. Personal experience has also shown that GRAIL2 may predict exons within repetitive elements. GRAIL2 searches both the forward and reverse strands and can also recognize translation start and stop signals. MZEF (Michael Zhang's Exon Finder; Zhang, 1997) predicts internal coding exons based on quadratic discriminate analysis. MZEF starts with a potential exon (an ORF 18 to

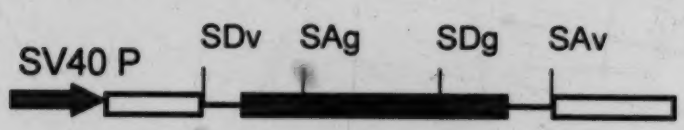
999 bp bounded by the correct splicing acceptor and donor sites), measures 9 discriminate variables and then calculates its exon probability. If the probability (P) is > 1/2, it will be output as a predicted exon. Personal experience has shown that the exons predicted by GENSCAN are most accurate and reliable, followed by MZEF and then GRAIL2.

The genomic sequence was also submitted to the BLAST series of searches following masking of the repetitive elements by RepeatMasker2 (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker). The BLAST hits obtained were compared to the exons and genes as well as previously identified exon and cDNAs in the region. In some cases ESTs were obtained for further characterization.

Figure 2-1: Exon Trapping System

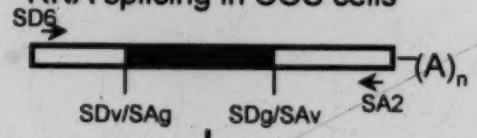
Exon trapping is a relatively simple approach for isolating transcribed sequences (exons) from cloned genomic DNA. The splicing vector pSPL3 contains two exons (open boxes) with an intron (thin horizontal line) from the HIV *tat* gene containing a vector splice donor site (SD_V) and a vector splice acceptor site (SA_V), as well as a SV40 promoter. The genomic DNA, subcloned into the multiple cloning site in the intron, may contain an exon (blue box) flanked by intron sequence (pink boxes). The genomic splicing signals are represented by SD_g and SA_g. When the cloned DNA contains an exon in the proper orientation, splicing can occur between the vector and insert sequences. A unique fragment is produced by subsequent PCR (primers indicated by small arrows).

Genomic DNA fragment subcloned into pSPL3



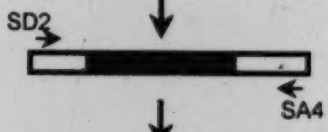
transfect into COS cells

transcription from SV40 promoter and RNA splicing in COS cells



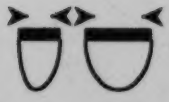
isolation of RNA and synthesis of cDNA by reverse transcription

1° PCR with SD6 and SA2



2° PCR with SD2 and SA4

Subclone into pAMP10



Colony PCR with SD2 and SA4

Run PCRs on 2 % agarose gel

Determine sequence of putative exons

FIGURE 2-2: A Detailed Flow Chart of Marathon cDNA Amplification Protocol

A) The gene-specific primer (GSP1 or GSP2, depending on type of RACE reaction) and AP1 are added to the Marathon double-stranded cDNA and other PCR components and thermal cycling is begun. In 5' RACE, the lower strand (antisense) cannot be extended or serve as a template because there is no AP1 binding site. Likewise in 3' RACE, the upper strand (sense) contains no AP1 binding site and is not extended.

B) Extension of the GSP in the first cycle of PCR will create AP1 binding sites on the lower and upper strands (5' and 3' RACEs, respectively).

C) Subsequent cycles of PCR allow extension from both the AP1 and gene-specific primers producing a 5' or 3' RACE product (D).

D) The 5' or 3' RACE product.

E) In some cases where the level of background or nonspecific amplification in the 5' or 3' RACE reaction is too high with a single gene-specific primer, then a nested or inner primer set may be used. A nested primer (AP2) is supplied with the kit and overlaps slightly with AP1. AP2 can be used with a nested gene-specific primer (NGSP1 or NGSP2) to reamplify an aliquot of the primary PCR.

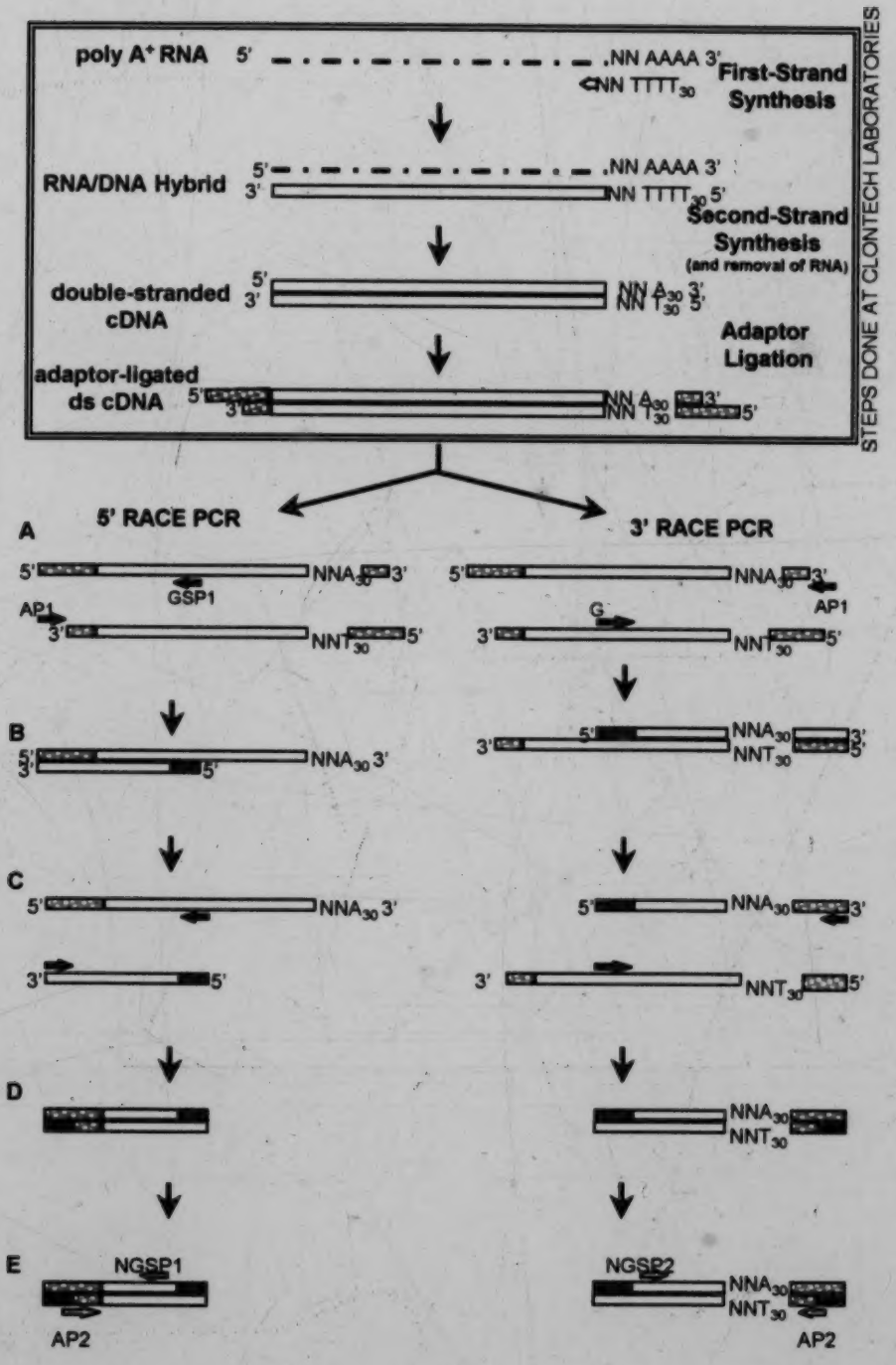
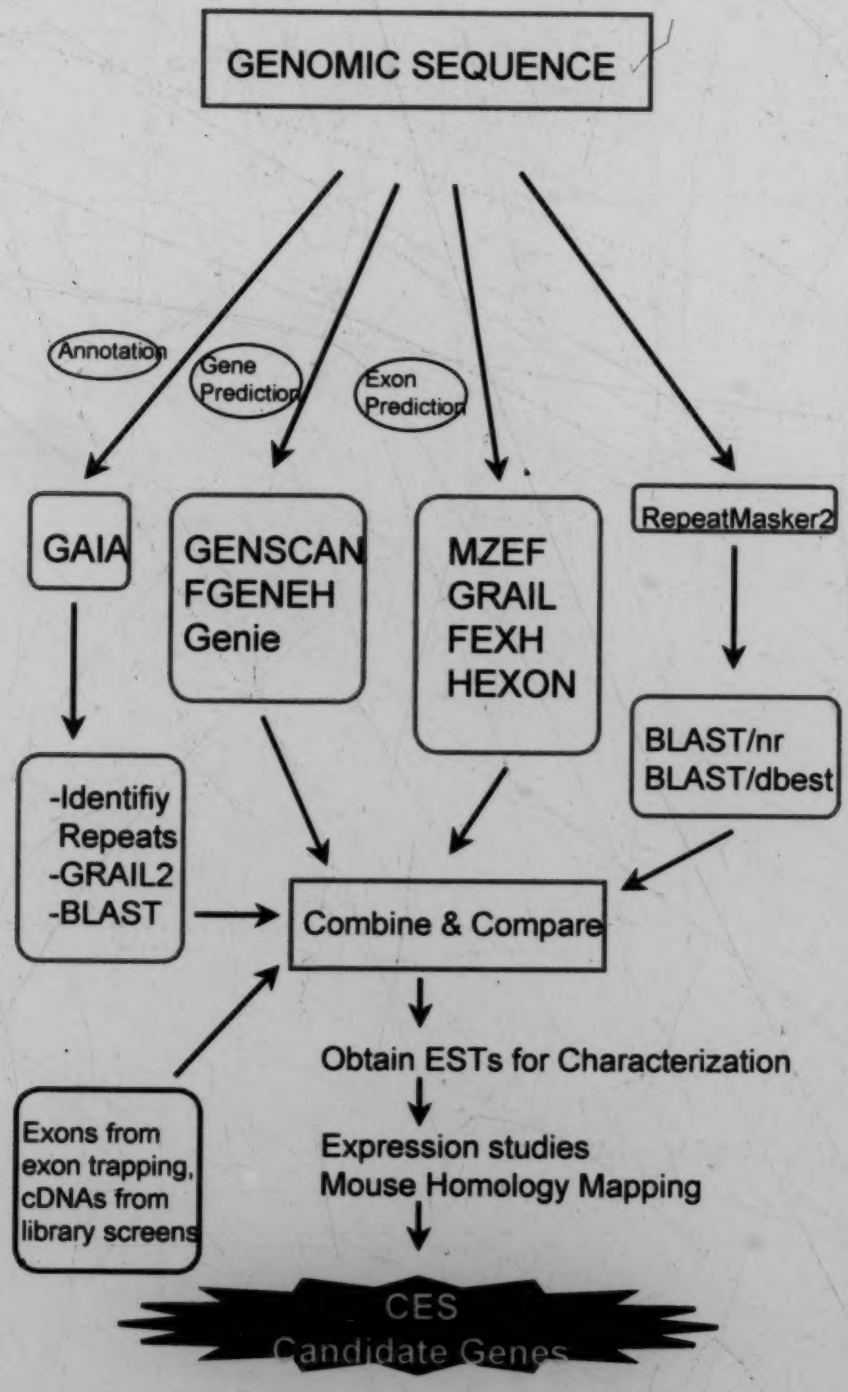

poly A+ RNA
5'
.NN AAAA 3'
<NN TTTT30
First-Strand Synthesis
RNA/DNA Hybrid
5'
3'
.NN AAAA 3'
NN TTTT30 5'
Second-Strand Synthesis
(and removal of RNA)
double-stranded cDNA
5'
3'
NN A30 3'
NN T30 5'
Adaptor Ligation
adaptor-ligated ds cDNA
5'
3'
NN A30
NN T30
3'
5'
STEPS DONE AT CLONTECH LABORATORIES
5' RACE PCR
3' RACE PCR
A
5'
NNA30 3'
GSP1
AP1
3'
NNT30 5'
G
AP1
B
5'
3'
NNA30 3'
5'
C
5'
NNA30 3'
3'
5'
NNT30 5'
D
NNA30
NNT30
E
NGSP1
AP2
NGSP2
NNA30
NNT30
AP2

Figure 2-3: Genomic Sequence Analysis Flow Chart

Genomic sequence was sometimes submitted to GAIA for formatting and annotation. Alternatively, the sequence was analyzed manually by submission to gene and exon prediction programs followed by masking of interspersed repetitive elements and BLAST searches. All of this information was compiled together with other previously known exons and cDNAs. In some cases, ESTs were obtained for further characterization such as sequencing, Northern blot analysis, RT-PCR, RACE etc..


GENOMIC SEQUENCE
Annotation
Gene Prediction
Exon Prediction
GAIA
GENSCAN
FGENEH
Genie
MZEF
GRAIL
FEXH
HEXON
RepeatMasker2
-Identifiy Repeats
-GRAIL2
-BLAST
BLAST/nr
BLAST/dbest
Combine & Compare
Obtain ESTs for Characterization
Exons from exon trapping, cDNAs from library screens
Expression studies
Mouse Homology Mapping
CES
Candidate Genes

TISSUE	PRIMING	VECTOR	SOURCE	MADE BY
Fetal Brain	Oligo d(T)	λZapII	Stratagene	Stratagene
CaCo-2 cell line	Oligo d(T) or Random Hexamers	λZap or pEXLOX	Dr. Joanna Rommens	Dr. Joanna Rommens
Adult Pancreas	Oligo d(T)	λZapII	Dr. Joanna Rommens	Dr. Joanna Rommens
Fetal Craniofacial	Oligo d(T)	λZapII	Dr. Michael Walter	Dr. Jeff Murray
Adult Skeletal Muscle	Oligo d(T)		Dr. Ali Riazi	Dr. R. Farahani
Adult Heart	Oligo d(T) & Random Hexamers	λZapII	Dr. Larry Fliegel	CLONTECH

Table 2-1: A List of the cDNA Libraries Used .

PUTATIVE GENE	PRIMER NAME	PRIMER 5'-3' SEQUENCE
CES86	EX86K1-1	ATGCACAGAGCACATATTCT
	EX86K1-1R	ACCTGAATACGAAAGAACAT
	EX86K1-2	TTCAAAGTCCAACACAGAAT
	EX86K1-2R	ACCTTTTCCACATTACATCA
	EX86-1	GTATAAAGCCCAGCCCTCCG
	EX86-2	AGAGAAGCTGGAAGAAAGAAC
	EX86-3	GACAGGCAGAACACAGACTT
	K1A-1	ATATAGCCTTATGACTGGCT
	K1A-2	GTACCGTCTGTTGTACAAAT
	K1A-3	GAACGTGTTTACACAGACGT
	EX86HEST-1	TCCGTTGCAAGGAAGAAGTG
	EX86HEST-2	CCGCGCAATGTTACCGGGAT
	HEST-3	TCACCGCGATCTCGGTCTAGG
	HEST4	TTTCTGTAAACCTGCTGACG
CES38	EX38Q1-1	TTTTGCTTTGGTTTCAGAAC
	EX38Q1-2	AACCTTGTCAGATAACTAAC
	EX38Q1-3	GAATATTAGGCAAGCTTGAT
	EX38Q1-4	GTATTTTAAAGCCAGATGGT
	EX38Q1-5	TTCTTCATGGAGTCCTAGAA
	EX38Q1-F1	ACCAATAAGTAACCTGTAGAGGTC
	EX38Q1-R1	CTGAGAGAGAGACAGAAGCAG

Table2-2: List of all the Primers Used

CHAPTER 3: RESULTS

A) Characterization of Putative Exons Within the CESCR

Exon Amplification

To isolate putative transcribed sequences from the CESCR, Dr. Ali Riazi performed exon amplification (see Fig. 2-1 for an overview of this technique) on 2 BACs (95A8, 233A2), 3 PACs (109L3, 238M15, 120N18) and 27 cosmid clones (Riazi, Ph.D. thesis, 1998). Clones subjected to exon amplification are starred on Figure 3-2.

Dr. Riazi digested these clones with *Sst* I and shotgun subcloned the fragments into the *Sst* I digested exon trapping plasmids pSPL3 or pSPL3B and plated them on LB + Ampicillin plates. He then screened colony lifted filters of these subclones with a probe made from total human genomic DNA. This was done to alleviate the problems of the amplification of fragments derived from the cosmid/BAC/PAC vector or other possible DNA contaminants. He found on average that approximately 60 % of the subclones hybridized to the total human DNA. After isolating DNA from these subclones he transfected them into COS cells, then isolated RNA and synthesized cDNA using reverse transcription. Dr. Riazi then performed primary and secondary PCR on a pool of cDNAs and subcloned the PCR products into the cloning vector pAMP10. Transformed cells were plated on LB + Ampicillin plates and single colonies re-grown on LB + Ampicillin plates in order to isolate single colonies. Dr. Riazi characterized a large number of clones, resulting in 9 putative exons. I analyzed approximately 119 remaining clones as the basis of this thesis.

Colony PCR

Each of the 119 clones were subjected to colony PCR using a single bacterial colony and the secondary amplification primers USD2 and USA4 (provided with Exon Trapping Kit). Approximately half of the PCR reaction was electrophoresed on a 2 % agarose gel in order to size the putative exons and to resolve small size differences between the various products.

A source of contamination in the Exon Trapping procedure was "vector-vector" splicing, which occurs when the vector splices the HIV tat exons together without an insert. When this occurs, secondary PCR will produce a 177 bp product. Alternatively, when a putative exon is inserted, secondary PCR will produce a product consisting of 177 bp (HIV exon sequence) + *X* bp of putative exon sequence. These size differences were easily discerned when the secondary PCR products were electrophoresed on a 2 % gel (See Figure 3-1). Figure 3-1 shows 19 colony PCR products. Lanes 2, 6, and 17 show bands approximately 177 bp in size, corresponding to vector-vector splicing. These clones were immediately eliminated from further characterization. Lanes 3, 4, 7, 9, 10, 13, and 18 show bands which are greater than 177 bp and likely contain a putative exon. All such clones were characterized further. Lane 5 shows three different sized bands and likely indicates a mixture of colonies was used in the colony PCR. Such a result was seen only 2-3 times and these clones were not characterized further. Lanes 1, 8, 11, 12, 14, 15, 16, and 19 show no product from the colony PCR. These results are theoretically inconsistent with the exon trapping procedure but may be due to contaminating ampicillin-resistant bacteria. Clones which showed such a result were re-grown on LB + ampicillin plates and the colony PCR was repeated if growth occurred.

Table 3-1 summarizes the results from all clones analyzed and lists the number of putative exons discovered. From BAC 95A8 there were 27 pAMP10 clones analyzed. Of these 27, 15 were due to vector-vector splicing, 10 showed no secondary PCR product and the remaining two were putative exons, which when sequenced turned out to be identical to two exons (Exons 7 and 8) already being characterized by Dr. Ali Riazi. From PAC 109L3 there were 13 pAMP10 clones analyzed. Of these 13, 5 were due to vector-vector splicing, 7 showed no secondary PCR product, leaving one putative exon. From PAC 238M15 there were 79 pAMP10 clones analyzed. Of these 79, 16 were due to vector-vector splicing, 30 showed no secondary PCR product and the remaining 33 were putative exons. Clones from BAC 233A2 and PAC 120N18 were characterized by Dr. Riazi.

Sequencing and BLAST Searches

Each of the putative exons was sequenced manually. Comparison of the sequences eliminated those clones which were identical. This resulted in 9 exons: Ex 11A, Ex 20, Ex 38, Ex 41, Ex 45, Ex 48, Ex 51, Ex 60 and Ex 86. Subsequently, the sequence of each exon was subjected to the BLASTN and TBLASTX programs to identify any similarities to previously known human or non-human genes or ESTs. Table 3-2 lists the sequence of each exon and summarizes the results of the BLAST searches. Ex 60 was found to be 100 % identical to a previously trapped exon (Troffater 1995) but identified no ESTs. Ex 86 identified one EST initially, EST 321686 (accession number W35386) from a parathyroid tumor which was obtained for further characterization. This EST is also referred to as hEST1. Several months later two other ESTs were also identified by Ex 86: EST 462605 (accession number A704966) from a

combined fetal liver and spleen cDNA library and EST 966077 (accession number AA505576) from breast tumor tissue. EST 462605 was obtained for further characterization and is referred to as hEST2. EST 966077 was not obtained because the information available regarding it's sequence and size indicated that it was identical to hEST1, extending only three bp more 5' than hEST1. Ex's 45 and 51 were found to contain repetitive DNA sequences; Ex 45 was entirely part of an Alu element and Ex 51 was part of a MER1 DNA element. Ex 48 was found to contain 81 bp of a LTR element while the remaining 64 bp were unique sequence. Ex's 11A, 20, 38, and 41 did not contain any repetitive DNA elements but did not identify any ESTs or genes.

ESTs identified by the BLAST search with Ex 86 sequence will be further discussed in section B.

Mapping Within the CESCR

Each putative exon was also hybridized to some of the PAC/BAC clones in the CESCR which were available at the time. This was done in order to map the exons to the smallest possible region within the CESCR. Since the physical map of the CESCR (Johnson et al., 1999) contained many overlapping PAC/BAC clones it was possible to map most exons to a region of 50 kb or less.

Figure 3-2 shows a partial physical map of the CESCR (derived and redrawn from Johnson et al. 1999) and indicates the regions where the exons mapped. The region shown by A contains Ex 86, which hybridized strongly to a single band in PAC 109L3. Clones 213P2, 609C6 and 143I13 were not available at the time for this hybridization experiment, however Ex 86 was later amplified by PCR from these three clones. The region shown by B contains Ex 20, 38 , and 51 which hybridized strongly to

clones 238M15 and 609C6. It is unclear why Ex 51 hybridized strongly to these clones since it is composed entirely of repetitive DNA. The region shown by C contains Ex 60 only, which hybridized strongly to only 238M15. The region indicated by D contains Exons 11A, 41, 45, and 48. Exons 11A, 41, and 48 hybridized strongly to single bands in 238M15 and 115F6 while Ex 45 hybridized strongly to a single band in 238M15 and 115F6 but also faintly to several other bands in other clones. This could be explained by the fact that Ex 45 is composed entirely of repetitive DNA.

Determination of Loci Number

As a preliminary experiment, each putative exon was hybridized to a partial hybrid panel to gain information regarding the number of loci in the genome which were represented by each exon. The pericentromeric region of chromosome 22 is known to be rich in low copy and interspersed repetitive DNA elements as well as truncated unprocessed pseudogenes including those for vonWillebrand factor (Eikenboom et al. 1991), neurofibromatosis (NF1; Regnier et al. 1997), adrenoleukodystrophy (ALD; Eichler et al. 1997), KCNMB3L (Riazi et al. submitted), the immunoglobulin kappa light-chain gene (IGKV3; Lotscher et al. 1986) plus many others (Minoshima et al., 1998; McDermid, unpublished results). Exons of unprocessed pseudogenes are equally as likely to be isolated by exon amplification as are exons of real genes provided their splice junctions are conserved.

Before extensive characterization of a putative exon was performed, it was hybridized to a Southern blot containing the following DNA digested with one or more restriction enzymes: total human DNA (cell line GM03657), DNA from a somatic cell hybrid (human/hamster) cell line containing chromosome 22 as the only human component (GM010888)

and total hamster DNA (cell line RJK888). The results obtained from this type of Southern blot may alert one to the possibility that an exon is part of an unprocessed pseudogene in the CESCR. It is important to note that this experiment may not identify all pseudogenes. If an exon represents 2 loci in the genome then one could reasonably expect to see a greater number of bands within total genomic DNA than in chromosome 22 only DNA, as a result of one functional ancestral locus probably located on another chromosome and a second locus (possibly a pseudogene) located within the CESCR on chromosome 22. Alternatively, the extra band(s) in total genomic DNA could be the result of a polymorphic restriction site within the probe being used. However, if a duplication was very recent (thus preserving all restriction sites), the band could be the same size on different chromosomes, requiring a complete monochromosomal hybrid panel to detect. Alternatively, both the ancestral gene and the duplicated gene fragment could be located on chromosome 22, in which case there would be no differences between the hybridization pattern of total genomic DNA and chromosome 22 only DNA. Any putative exons that hybridized to more bands in total human DNA than in chromosome 22 only DNA, and therefore may have been part of a pseudogene within the CESCR, were not characterized any further. Table 3-3 shows the results of these experiments.

Exons 11A, 20, 38, 41, and 86 hybridized to single bands in both the total human DNA and the chromosome 22 only DNA and likely represent single loci. Ex 48 hybridized to two bands in the total human DNA, only one of which hybridized in chromosome 22 only DNA. It is therefore possible that this exon represents more than one locus in the genome. Exons 45 and 51 produced a smear, which is consistent with the fact that they are both entirely repetitive DNA. No result was seen with Ex 60. This may be due to the small size (50 bp) of this exon and the

limitations of hybridization. Exons 11A, 20, 38, 41, and 86 which hybridized well, were all substantially larger (93-201 bp) than Ex 60. Exons 45, 48, and 51 were not subjected to further characterization.

Screening cDNA Libraries

Exons 11A, 20, 38, 41, 60, and 86 were used to probe a CaCo-2 cDNA library and a fetal brain cDNA library (see table 2-1 for a description of these cDNA libraries). The CaCo-2 library was produced in the lab of Dr. Joanna Rommens from RNA isolated from a colon carcinoma cell line. This cell line is known to express many widespread and tissue-specific genes including the CFTR gene responsible for cystic fibrosis (Rommens, personal communication), uroplakin (a bladder-specific gene) and ARHGAPL (a gene expressed mainly in kidney and placenta) (McDermid, unpublished results).

Screening of the fetal brain library produced two cDNA clones; one containing Ex 86 (called Ex 86 K1) and one containing Ex 38 (called Ex 38 Q1). No cDNAs were isolated for Exons 11A, 20, 41, or 60. These exons may be contained within genes which are not expressed in the fetal brain or CaCo cell line or at the developmental age of the fetal brain used to make this library. Alternatively, they may be exons located at the extreme 5' end of a gene and therefore be underrepresented in the libraries tested. The two cDNAs obtained will be discussed in Section B.

Overall two putative genes were discovered from nine initial exons:CES38 (CES putative gene containing Ex 38) and CES86 (CES putative gene containing Ex 86). The characterization of these putative genes is described in section B.

B) Characterization of Putative Genes in the CESCR

1) Partial Cloning and Characterization of a Putative Gene, CES86

Mapping cDNA Clones Within the CESCR

The sequence of Ex86 (Table 3-2) originally identified one EST, EST 321686 (referred to as hEST1;Accession # W35386), and later on two other ESTs, EST 462605 (referred to hEST2; Accession # AA704966) and 966077 (Accession # AA505576). hEST1 and hEST2 were obtained for further characterization and sequencing.

Screening of fetal brain and CaCo-2 cDNA libraries with Ex86 resulted in a 5 kb cDNA clone (Ex86 K1) from the fetal brain library which was further characterized.

After sequencing of approximately 350 bp at the ends of the Ex86 K1 cDNA, PCR primers were designed and used to amplify these sequences from the PAC/BAC clones from the CESCR. The first primer pair, K1-1 and K1-1R, at the T7 end of the insert (See Figure 3-3) amplified a 200 bp fragment from clones 109L3, 213P2, 143I13, and 609C6. The second primer pair, K1-2 and K1-2R, at the T3 end of the insert did not amplify an expected 200 bp fragment from the four previously mentioned clones. Since the exon-intron boundaries were unknown, it was possible that the second set of primers extended across a large intron that would not be amplifiable by PCR. To circumvent this problem, the cDNA was restriction mapped and various restriction fragments were hybridized to many of the PAC/BACs in the CESCR (including 916F2, 109L3, 143I13, 213P2, 609C6, 238M15, 567H5, 50A16). As seen in Figure 3-3, four contiguous fragments did not hybridize to PAC/BACs: *Eco*RI/*Bam*HI 1.1, *Bam*HI/*Hind*III 0.1, *Hind*III/*Xba*I 0.4 and

*Xba*I/*Eco*RI 0.1. The *Eco*RI/*Pst*I 0.55 fragment was not tested because it could not be easily isolated as a probe. The rest of the fragments tested did hybridize to CESCR PAC/BAC clones. Together, these experiments suggested that the K1 cDNA was chimeric. The chromosome 22-specific portion of the K1 cDNA, approximately 2.6 kb in size, was subcloned using the internal *Hind*III site and the *Hind*III site located in the multiple cloning site of the vector at the T7 end. The subcloned portion was referred to as Ex86 K1a and was subjected to further characterization and sequencing.

Sequence Analysis of Three cDNAs: Ex86 K1a, hEST1 and hEST2

The analysis of the sequence obtained for K1a showed no open reading frame (ORF) as well as the presence of portions of several repetitive elements including LTR, L2, Alu, and MER1 (See Figure 3-4). Sequencing of Ex86 K1 using the primer 86-2, gave 106 bp of sequence toward the T3 end, past the *Hind*III site used for subcloning. Analysis of this sequence (see Figure 3-5) showed a consensus polyadenylation signal but no poly (A) tail.

Sequencing of hEST1 and hEST2 revealed that they both contained a polyadenylation signal and poly(A) tail and therefore represented the 3' end of this putative gene (See figure 3-6). Ex86 K1a and hEST2 each contained the original trapped Ex86 in its entirety. hEST1 was found to contain only the 3' portion (70 bp) of Ex86 spliced to sequence not present in Ex86 K1a or hEST2. Therefore Ex86 was subdivided into two portions called 86-A (3' 70 bp) and 86-B (5' 57 bp).

This finding led to the hypothesis that the trapped Ex 86 was actually two separate exons which had been trapped together. However, Ex 86 (127 bp) could be amplified by PCR using the primers 86-1 (or 86-

3) and 86-2 (See Fig. 3-4) from four overlapping PACs and BACs (PACs 109L3, 143I13, 213P2 and BAC 609C6) and from total human DNA, indicating that it was contiguous in the genomic sequence and therefore not two separate exons trapped together (results not shown).

The sequence of Ex86 K1a could also be amplified by PCR using the primers 86-1 (or 86-3) and K1a-2 from the same four overlapping PAC/BACs. This implied that the entire Ex86 K1a cDNA was one large 3' exon or that it was created from priming off genomic DNA and not mRNA when the cDNA library was made. The idea that Ex86 K1a was made by priming off genomic DNA was also supported by the presence of a correct 3' splice acceptor site and a 5' splice donor site at the boundaries of Ex86 (See figure 3-5). However, the sequence of the conserved splicing signals in Ex86 K1a suggested that Ex86 was in the opposite orientation as hEST1 and hEST2. This same sequence however was also present in hEST2 which is polyadenylated and in the same direction as hEST1. The correct 3' and 5' splicing signals around Ex86 may be cryptic splicing signals which allowed it to be trapped by the exon trapping procedure. In any case this meant that the sequence unique to hEST1 (500 bp at the 5' end) must be located more 5' (of the sequence on Ex86 K1a and hEST2) in this putative gene. This 5' exon had previously been amplified by PCR using the primers H1 and H2 from the same four clones mentioned above. A portion of this exon does contain an open reading frame of approximately 326 bp (108 amino acids) which shows no similarity to previously identified genes or ESTs.

Complete genomic sequence of PAC 143I13 became available some time later and has confirmed all these PCR experiments to be correct as well as indicating the direction of this putative gene to be centromere to telomere (5' to 3'). Genomic sequence of PAC 143I13 showed that the distance of the intron in hEST1 was 5203 bp. As well, the

genomic sequence was analyzed for splice sites around hEST1. Correct splice sites were observed for both the 3' splice acceptor and the 5' splice donor.

Expression Analysis of CES86

To study the expression profile of this putative gene, a probe was made from the 5' exon of hEST1 (*Pst*/*Eco*RI 0.5 fragment) and hybridized to Clontech and Invitrogen Northern blots containing mRNA from adult or fetal tissues. Hybridization revealed a relatively abundant 7 kb transcript in adult heart and skeletal muscle and fetal heart and muscle (See Figure 3-7). A transcript approximately 7.8 kb in size was detected in all fetal tissues tested except heart, where the transcript was slightly smaller at approximately 6.1 kb. As well, a fainter transcript 2.1 kb in size was detected in several tissues.

A probe was also prepared from the entire hEST2 (*Pac*I/*Eco*RI 0.4 +0.7) and hybridized to commercially available blots. Hybridization revealed a 7 kb transcript in most adult and fetal tissues tested (See Figure 3-8).

To confirm the validity of Ex86 K1a and hEST1 cDNAs RT-PCR experiments were performed. Using total RNA isolated from various human adult and fetal tissues as well as total RNA isolated from 2 cell lines, reverse transcription was performed using the 86-3 primer. Subsequent PCR with 86-3 and H1, followed by a nested PCR with primers 86-3 and H4, produced a smear on an agarose gel, but when Southern blotted and probed with hEST1 *Pst I*/ *Eco* RI 0.5 showed the expected sized band of 150 bp in RT reactions from skeletal muscle, tonsil, thymus, heart, fetal brain and CaCo. No amplification was detected in the negative RT reactions (See Figure 3-9). This indicates that the

hEST1 cDNA is a real transcript in the tissues tested, however it may be relatively rare since the expected sized RT product could only be visualized with Southern hybridization. PCR using the primers 86-3 and K1a-2 produced a clearly visible band approximately 2.2 kb in all tissues tested, however the same band was also seen in negative RT controls indicating possible genomic DNA contamination. This result was inconclusive as the band produced in the RT reaction could have been amplified either from a true mRNA or from genomic DNA (results not shown).

5' End Cloning of hEST1

To identify more of the 5' end of this putative gene, hEST1 *PstI*/*Eco*RI 0.5 was used to probe several more cDNA libraries including the fetal brain library, as well as adult heart, adult skeletal muscle and adult pancreas libraries, a fetal craniofacial library and a CaCo-2 randomly primed library. No cDNA clones were obtained from any of these libraries. This is surprising especially considering the high expression of hEST1 in adult heart and skeletal muscle (as seen in Northern analysis).

To clone the 5' end of CES86, 5' RACE was also performed from hEST1. Three different Marathon RACE kits were used for this purpose: fetal brain, adult brain and adult heart (For an overview of the Marathon RACE procedure see Fig. 2-2). A primary PCR was performed with the primers H2 and AP1 (supplied with kit), followed by a nested PCR with H3 and AP2 (also supplied with kit). The RACEs from the fetal and adult brain kits were unsuccessful and gave no further 5' sequence. This could have been due to low or no expression of the transcript associated with the 5' exon of hEST1 in fetal or adult brain. The RACE from the adult heart kit yielded a product containing 5 more base pairs of 5' sequence that had

previously been unknown. This RACE product appeared to be chimeric in that the new 5 bp of 5' sequence was followed by a string of approximately 20 T's followed by sequence from a previously identified and well characterized gene (ATP synthase E chain from human, cow, rat, and mouse) presumably on another chromosome. Repeated RACE reactions did not produce any more 5' sequence.

3' End Cloning of hEST1

To obtain more information regarding the 3' end of CES86, 3' RACE was performed from hEST1 using the three Marathon RACE kits previously mentioned. A primary PCR was performed with the primers H1 and AP1, followed by a nested PCR with H4 and AP2. The RACEs from the fetal brain and adult heart kits were unsuccessful and gave no further 3' sequence. The RACE from the adult brain kit yielded two successful products (see Figure 3-4). 3' RACE product #1 contained 49 bp of the 5' exon, splicing (with correct 5' splice donor) before that of hEST1, to the 3' portion of the cDNA. 3' RACE product #2 contained the 104 bp 3' of the nested H4 primer and then continued on for 812 more bp contiguous with the genomic sequence. This RACE product was likely amplified from contaminating genomic DNA within the RACE kit used. This RACE product was also likely chimeric, since it then contained sequence not present on PAC 143I13, PAC 238M15 or PAC 109L3, which showed no sequence similarity with any other genes or ESTs.

Southern Analysis of CES86

As presented in Table 3-3, Southern analysis using Ex 86 as a probe suggested that it represented only one locus in the genome (See Figure 3-10). Further analysis with hEST1 was later performed using the *PstI*/*Eco*RI 0.5 fragment to probe a genomic Southern blot containing DNAs digested with the restriction enzymes *Sst* I, *Pst* I, *Hin*dIII and *Taq* I. Single bands were detected in total genomic DNA and chromosome 22-only DNA digested with *Sst*I and *Hin*dIII (results not shown). However, when these DNAs were digested with *Pst*I or *Taq*I, two bands were seen in total genomic DNA only one of which was seen in chromosome 22-only DNA. These results could have been explained by several possible theories:

1) The CES86 gene is a single-copy gene located within the CESCR and it contains *Pst* I and *Taq* I restriction sites which are polymorphic. Since the human cell line tested shows two bands, it is a heterozygote whereas the chromosome 22-only cell line is a hemizygote and therefore can only demonstrate one band.
2) The CES86 gene is a pseudogene which was duplicated from an ancestral locus located elsewhere in the genome. The *Sst* I and *Hin*dII restriction recognition sites were conserved between the original gene and this pseudogene, thus showing only one hybridizing band. The *Pst* I and *Taq* I restriction sites may or may not be conserved or new sites may have been created by changes in the DNA sequence, leading to two hybridizing bands in total human DNA, only one of which is represented on chromosome 22.
3) The CES86 gene is a pseudogene duplicated from elsewhere on chromosome 22, and producing the same sized restriction fragments, and one locus shows restriction site polymorphisms.

To test hypothesis #1, the *Pst*I/*Eco*RI 0.5 fragment was used to probe several Southern blots containing genomic DNA (digested with *Pst* I or *Taq* I) from various CES patients and their parents. If this theory was correct, then there would be some proportion of people carrying one or the other of the alleles (homozygous) and people carrying both alleles (heterozygous). This theory was supported by several different individuals (See Figure 3-11). This figure shows a Southern blot containing total genomic DNA, digested with *Pst*I, from 11 different individuals. When probed with hEST1 *Pst*I/*Eco*RI 0.5, four out of the eleven individuals showed only one allele, and were therefore homozygous. The remaining seven individuals showed both alleles and were therefore heterozygous. This result indicates that the CES86 gene contains a polymorphic *Pst*I restriction site. Therefore the two bands originally detected in total human DNA were due to this restriction fragment length polymorphism (RFLP) and not a second locus.

CES86 was also found to contain a polymorphic *Taq* I restriction site. The hEST1 *Pst*I/*Eco*RI 0.5 fragment was used to probe a Southern blot containing total genomic DNA from various individuals digested with *Taq* I (See Figure 3-13). As is seen in this figure, two out of six individuals show both the larger allele and the smaller allele and are therefore heterozygous. The remaining four out of six individuals show only one of the bands and are therefore homozygous.

To test hypothesis #2, the *Pst* I/*Eco* RI 0.5 fragment was used to probe two complete somatic cell hybrid panels (one with DNA digested with *Sst* I and another with DNA digested with *Hind* III). If this theory was correct, then the result would show single bands of the same size in total human DNA, chromosome 22 DNA and DNA from at least one other chromosome. The results obtained from six separate hybridization experiments were inconclusive (the bands were too faint to see even after

approximately two weeks of exposure to X-ray film). Satisfactory results may be obtained if a new hybrid panel is prepared with greater concentrations of DNA.

To further test hypothesis #2, the *Pst* I/*Eco* RI 0.5 fragment was used to probe several Southern blots containing total human DNA, chromosome 22 only DNA, and total hamster DNA digested with several different restriction enzymes (*Bgl* II, *Bsr* GI, *Eco* RV, *Hind* III, *Pvu* II, *Ssp* I, and *Sst* I). If CES 86 was a duplicated gene fragment from elsewhere in the genome then a discrepancy between the number and/or size of bands detected in total human and 22 only DNA would be expected. A number or restriction enzymes were used to cover a region of approximately 10700 bp surrounding the probe. This region extended distally into the 3' region of the CES86 cDNAs, and proximally into the flanking gene, PSL (see Figure 3-12). This figure shows the sizes and relative positions of the expected bands for each enzyme tested. For each enzyme, only the expected size band was detected in both total human DNA and chromosome 22 only DNA, suggesting that this region is not duplicated elsewhere in the genome.

CES86 Polymorphism Can Be Used to Determine Parent of Origin of CES Chromosomes

To determine parent of origin we may use one of several types of polymorphic probes including: Restriction fragment length polymorphism (RFLP), minisatellites (or variable number of tandem repeats; VNTR), or microsatellites. Affected individuals with a CES chromosome will harbour four copies of the CESCR; one from one parent and three from the other parent.

hEST1 *PstI*/*Eco*RI 0.5 was used to probe a Southern blot containing total genomic DNA, digested with *Taq I*, from several CES patients and their parents (See Figure 3-13). This Southern blot shows two families, one for which this probe was uninformative and one for which it was informative. In family two, the child (2C) inherits one copy of the smaller allele from the father (2F), and three copies of the larger allele from the mother (2M). The CES86 polymorphism will likely also be useful for determination of parent of origin for many other instances of CES.

Genomic Sequence Analysis of PAC 143I13

Sequencing of a number of overlapping PAC/BAC clones within the CESCR has been started by our collaborator Dr. Bruce Roe at the University of Oklahoma. The partial sequence of PAC 143I13 was available in the fall of 1998. The complete sequence of this PAC has been available since early 1999. This PAC clone contains the three Ex86 cDNA clones which were characterized in this study. The genomic sequence was analyzed for the presence of computer predicted genes and exons (in a collaborative effort by several individuals including myself, Graham Banting, Tim Footz, Stephanie Maier and Dr. Ali Riazi) in order to find the rest of the CES86 gene and it's location with respect to other genes (See Figure 3-14). Several putative genes were identified and ESTs were obtained for further characterization and comparison with the putative CES86. It was therefore necessary to partially characterize the flanking genes to see if CES86 was a part of a neighboring transcript.

2) Putative Genes Flanking CES86

Analysis of the genomic sequence of PAC 143I13 indicated that CES86 was flanked distally by IDGFL, a gene previously identified and being characterized by Dr. Ali Riazi and proximally by several EST clusters representing up to 5 genes.

A) An Insect-Derived Growth Factor-Like Gene (IDGFL)

This putative gene was identified through exon trapping by Dr. Ali. Riazi. A BLASTN/dbest search on April 8 1999 using PAC 143I13, identified approximately 10 ESTs at the 3' end of this putative gene. A near full-length cDNA clone, EST 54445 (Accession # AA348024) had been previously obtained and characterized by Dr. Ali Riazi. Analysis of this cDNA sequence by Dr. Riazi identified an ORF of about 1.5 kb encoding a protein with similarity (34.8%) to a putative growth factor from *Sarcophaga peregrina* (IDGF, Homma et al. 1996), as well as to atrial gland granule-specific antigen (AGSA at 26.7 %) of *Aplysia californica* (Sossin et al. 1989)(Riazi, Ph.D. thesis, 1998). Southern hybridization of EST54445 by Dr. Riazi and analysis of the genomic sequence showed that the orientation of IDGFL is telomere to centromere (5' to 3').

The expression profile of IDGFL was studied by Northern analysis by Dr. Riazi. He identified a transcript approximately 3.5 kb in size found predominantly in adult lung and placenta and several fetal tissues (results not shown).

It is unlikely that CES86 is part of the IDGFL gene for the following reasons:

1) IDGFL is in the opposite orientation to CES86
2) The patterns obtained from Northern blots are non-overlapping.

3) EST 54445 is nearly a full-length cDNA clone (based on the Northern Hybridization experiment) and therefore most of the gene has already been cloned.

B) A Phosphatidyl Synthase-Like Gene (PSL)

This putative gene was identified through a collaborative genomic sequence analysis (gene and exon prediction programs) by several individuals in the lab. A BLASTN/dbest search on April 8 1999, using PAC 143i13 identified approximately 20 ESTs at the putative 3' end of this gene. The largest EST, EST 52444 (Accession # H23285), had previously been obtained for further characterization.

Analysis of the genomic sequence containing this putative gene indicated that its direction was telomere to centromere (5' to 3'). Splicing together of the predicted exons yielded an ORF of approximately 190 amino acids which shows similarity to a *S. cerevisiae* hypothetical 39.4 Kda protein, a *B. subtilis* arabinose operon protein, and *S. pombe* phosphatidyl synthase.

To study the expression profile of this putative gene a probe was made from the 52444 EST (an 800 bp *Pst*I fragment). This probe was hybridized to Clontech Northern blots containing mRNA from various adult and fetal tissues. Hybridization revealed an abundant 1.9 kb ranscript in all tissues tested (See Figure 3-15).

It is unlikely that CES86 is part of the PSL gene for the following reason:

1) PSL is in the opposite orientation as CES86

C) A Putative Single Exon Gene, BTPUTR

This putative gene was first identified through a collaborative genomic sequence analysis by several individuals in the lab. A BLASTN/dbest search on April 8 1999 using PAC 143i13, identified a cluster of approximately 20 ESTs spanning a total distance of approximately 2500 bp at the 3' end of this putative gene. The largest EST, EST 46414 (Accession # H09166) had previously been obtained for further characterization. This putative gene was called BTPUTR, standing for big three prime UTR.

Analysis of the genomic sequence containing this putative gene indicated that its direction is telomere to centromere (5' to 3'). Analysis of a large segment of contiguous and repetitive element-free genomic sequence, encompassing the predicted exons and EST cluster, yields two ORFs. The longer ORF of 292 amino acids shows no similarity to anything in the database. The shorter ORF of 268 amino acids contains predicted leucine zipper and helix-loop-helix motifs, both important in transcriptional regulation. Since the two putative ORFs along with the cluster of ESTs are contiguous in the genomic sequence, it is possible that this putative gene is composed of only one exon.

To study the expression profile of this putative gene a probe was made from the 46414 EST (the entire insert was excised using *Hind*III and *Not*I). This probe was hybridized to Clontech Northern blots containing mRNA from various adult and fetal tissues. Hybridization revealed strong expression of a 5 kb transcript in heart, brain, prostate, testes, peripheral blood leukocytes, and fetal brain. Weaker expression of the same transcript was detected in all other tissues tested (See Figure 3-16).

It is unlikely that CES86 is part of the BTPUTR gene for the following reasons:

1) BTPUTR is in the opposite orientation as CES86
2) The patterns obtained from Northern blots are non-overlapping.

D) The Interleukin-17 Receptor Gene

The IL-17 receptor gene was previously mapped to 22q11.2 by (Yao et al., 1997). Genomic sequence analysis of PAC 143I13 and PAC 109L3 allowed the precise localization of this gene to the central region of the distal CESCR.

To study the expression profile of this gene, S. Maier prepared a PCR probe from the 3' end of this gene. This probe was hybridized to adult Clontech Northern blots. Hybridization revealed a complex pattern of multiple-sized transcripts (See Figure 3-17). Transcripts approximately 8.8, 6.3, 5.0, 2.6, 1.9, and 1.05 kb in size were detected in multiple tissues.

It is unlikely that CES86 is part of the Il-17R gene for the following reason:

1) The patterns obtained from Northern blots are non-overlapping.

Together, these experiments suggest that the CES86 gene is not likely to be part of the putative genes IDGFL, PSL, BTPUTR, or IL-17R. With this in mind, and given it's orientation (centromere to telomere), if CES86 is a functional gene within the CESCR then it's 5' end may extend off PAC 143I13 and into the more proximal PAC 109L3. If this were true, CES86 would have to contain at least three other genes within one or more of its introns. Such cases of nested genes is rare but has been documented. For example, nested within intron 26 of the NF1 gene are three genes (OGMP, EV12B, and EV12A) transcribed from the opposite DNA strand as NF1 (Viskochil et al. 1991). A similar situation exists in

intron one of the HIRA gene in the DiGeorge critical region. A multi-exon gene (22k48) is transcribed from the opposite DNA strand as HIRA (Pizzuti et al. 1999). As well, within intron 22 of the factor VIII gene there is a CpG island which promotes the transcription of two genes. One gene, F8A, is transcribed in the opposite direction as the factor VIII gene, while the second gene, F8B, is transcribed in the same direction as factor VIII. F8B contains a novel 5' exon from intron 22 sequence and then splices to the remaining four exons of the factor VIII gene (Levinson et al. 1992).

Therefore it was necessary to analyze the complete genomic sequence of PAC 109L3. The partial sequence of PAC 109L3 was available in June 1999. The genomic sequence was analyzed for the presence of computer predicted genes and exons in order to find the rest of the CES86 gene.

Genomic Sequence Analysis of PAC 109L3

The partial sequence of PAC 109L3 was available in the spring of 1999. The complete genomic sequence was analyzed for the presence of computer predicted genes and exons (in a collaborative effort by several individuals including myself, Graham Banting, Tim Footz and Stephanie Maier) in order to find the rest of the CES86 gene and it's location with respect to other genes (See Figure 3-18).

The genomic sequence was submitted to the GENSCAN gene prediction program, and the GRAIL2 and MZEF exon prediction programs, with and without masking of repetitive elements. The genomic sequence was also submitted to BLASTN/nr and BLASTN/dbest searches to find any ESTs or fragments mapping to other sequenced regions of the genome. All analysis of PAC 109L3 was performed on August 28, 1999. The proximal end of this clone showed interspersed

regions with similarities to chromosomes X, 21, 16, and 12. At least two putative pseudogenes were evident in this region (Gab2 and IL-9R). The central portion was found to contain a putative gene already characterized by Dr. Ali Riazi (SAHL). The distal portion was found to contain the complete coding sequence for the IL-17R gene, the putative genes BTPUTR and PSL. The exons of these putative genes were overall, well predicted by all three of the prediction programs used. A number of high scoring exon predictions by GRAIL2 were found proximal to, and on the same DNA strand as the CES86 gene. This finding suggests that further 5' exons of CES86 may be overlapping with PSL and/or BTPUTR, and transcribed in the opposite direction and from the opposite DNA strand as these two genes. There were no exons predicted proximal to IL-17R which would reasonably be a part of the CES86 gene.

3) Partial Cloning and Characterization of a Putative Gene, CES38

Mapping the Ex 38 cDNA Clone Within the CESCR

The sequence of Ex 38 (Table 3-2) identified no ESTs by performing BLASTN/dbest and TBLASTX searches. Screening of fetal brain and CaCo-2 cDNA libraries resulted in a ~ 2.3 kb cDNA clone (Ex38 Q1) from the fetal brain library which was further characterized. This putative gene is referred to as the CES gene containing Exon 38, CES38.

The Q1 cDNA obtained from the fetal brain cDNA library was restriction mapped, the cDNA was cut with various restriction enzymes and each fragment was hybridized to many of the PAC/BACs in the CESCR. Two fragments at the T7 end (*Eco*RI 0.4 and *Eco*RI/*Hind*III 0.26) did not hybridize to the PAC/BAC clones (See Figure 3-19). These hybridization results suggested that the Q1 cDNA was chimeric. It was

also known that the frequency of chimeric cDNAs in the fetal brain library was very high (this thesis; Riazi, 1998; A. Wong, personal communication). It was later determined through analysis of genomic sequence of PAC 238M15 that 528 bp of contiguous sequence at the T3 end of Q1 was in fact located on this PAC. The remaining 1679 bp were not located on this PAC or any other available genomic sequence. Using the chimeric portion, BLASTN/nr and BLASTN/dbest searches on July 11 1999, did not reveal any matching human or non-human genes or ESTs.

Sequence Analysis of the Ex38 Q1 cDNA

The Q1 cDNA was fully sequenced (See Figure 3-20) and analyzed. The sequence showed an ORF of 55 amino acids (165 bp) at the 5' end as well as the presence of an L1 repetitive element within the chimeric portion (See Figure 3-19 for a pictorial representation of the Ex 38 cDNA). As well, this cDNA did not contain a polyadenylation signal or a poly (A) tail. It therefore likely represents an internal gene fragment near the 3' untranslated region (UTR) of this putative gene. Analysis of the genomic sequence of PAC 238M15 and identification of conserved splicing sites around Ex38, revealed the direction of this putative gene as telomere to centromere.

Expression Analysis of the Ex 38 Q1 cDNA

To study the expression profile of this putative gene, a probe was first made from the entire cDNA (*Eco* RI fragments 1.8 + 0.4) and hybridized to a Northern blot containing mRNA from various mouse tissues and mRNA from the cell lines CaCo (colon carcinoma) and HeLa (cervical carcinoma). No transcripts were detected on this Northern blot. A

second probe was made which did not include the L1 repetitive element (*Eco* RI 1.9) and hybridized to Clontech Northern blots containing mRNA from adult or fetal tissues. No transcripts were detected on these Northern blots. Since no results were obtained using these two DNA probes, it was thought that this putative gene may be expressed at extremely low levels, which would not be detectable by Northern analysis, or in tissues not present on these Northern blots.

It was decided that since the Q1 cDNA was likely chimeric to an unknown degree and since no Northern results could be obtained that the characterization of this putative CES38 gene would be halted at least until genomic sequence for this region was available.

Work in the lab after experiments for this thesis had finished has revealed a successful Northern hybridization using the Ex38 Q1 F1/R1 PCR product (composed of the non-chimeric portion of Ex38 Q1) as a probe. The probe was hybridized to a Clontech Northern blot containing mRNA from various adult tissues. This hybridization showed a strongly expressed 2.4 kb transcript in lung as well as a weaker 3.7 kb transcript in all tissues tested (See Figure 3-21).

Southern Analysis of CES38

As presented in Table 3-3, Southern analysis using Ex 38 as a probe suggested that it represented only one locus in the genome. Further analysis with the non-chimeric portion of Ex38 Q1 was later performed, after experiments for this thesis had finished, using the Ex38 Q1 F1/R1 PCR product to probe a partial hybrid panel. This Southern blot contained total human DNA, chromosome 22 only DNA and hamster genomic DNA digested with the restriction enzymes *Bam*HI, *Sst* I, and *Taq* I. Hybridization showed single bands in total genomic DNA and

chromosome 22 only DNA digested with *Bam*HI or *Sst* I and two bands in those DNAs digested with *Taq*I (See Figure 3-22). Since the bands observed in chromosome 22 only DNA were identical in size and number to the bands observed in total genomic DNA, it is reasonable to assume that the PCR probe used for hybridization likely represents only one locus in the genome.

Genomic Sequence Analysis of PAC 238M15

The partial genomic sequence of PAC 238M15 was available in spring 1999. Hybridization experiments (previously discussed) showed that this clone contained Ex 38 and the Q1 cDNA as well as Exons 11A, 20, 41, 48, 60. The genomic sequence was analyzed for the presence of computer predicted genes and exons in order to find the rest of the CES38 gene and other putative gene(s) containing the trapped exons. The genomic sequence was submitted to the GENSCAN gene prediction program, and the GRAIL2 and MZEF exon prediction programs, with and without masking of repetitive elements. Several low or moderate scoring exons were predicted on both strands, however, no obvious genes could be predicted using GENSCAN (See Figure 3-23). The genomic sequence was also submitted to BLASTN/nr and BLASTN/dbest searches to find any ESTs or fragments mapping to other sequenced regions of the genome. All analysis of PAC 238M15 was performed on May 24, 1999.

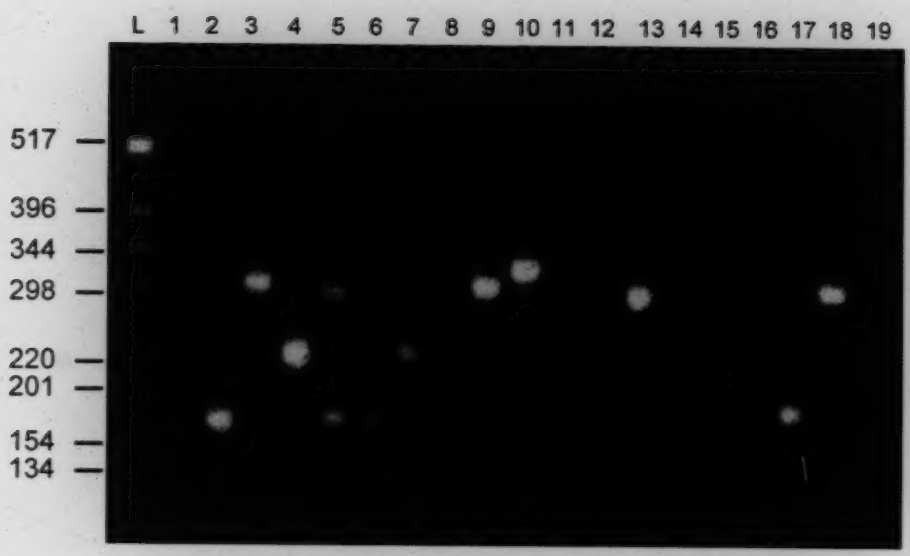
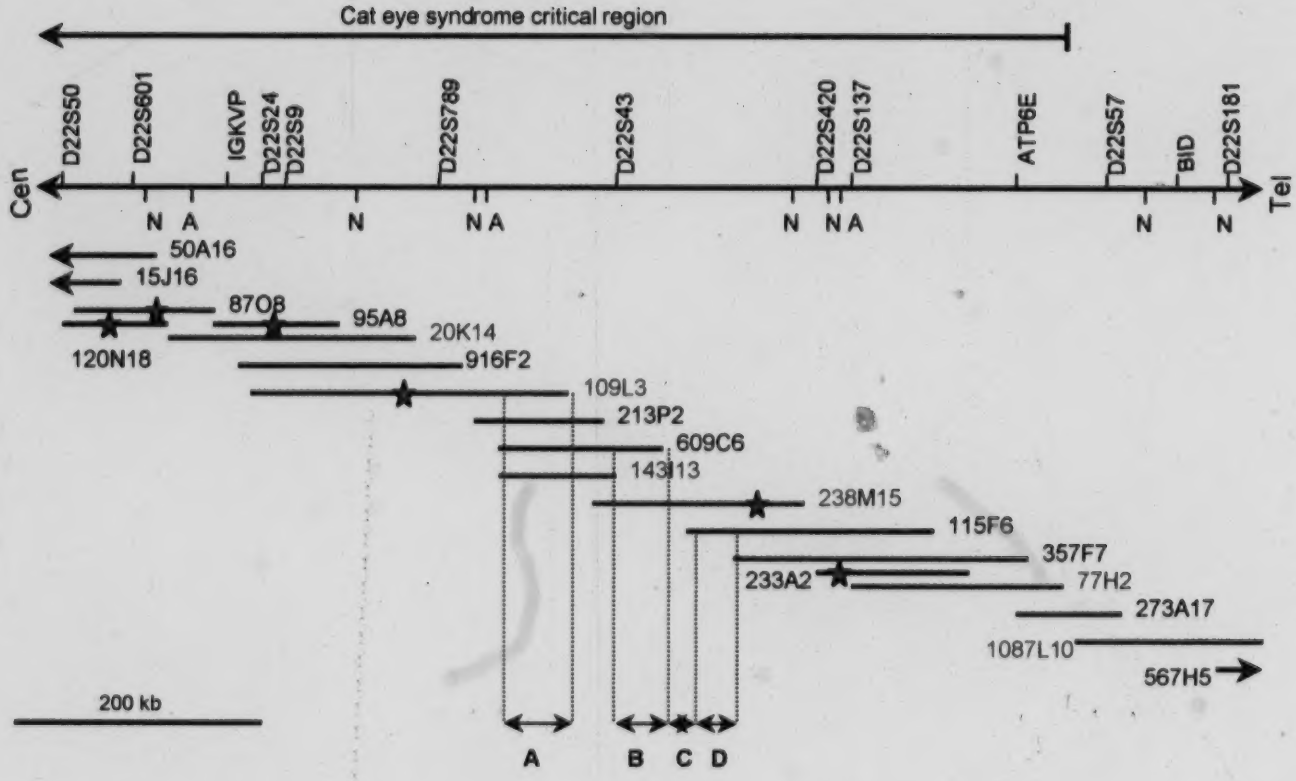


Figure 3-1: Colony PCR Products

Lane L contains a 1kb DNA size ladder with sizes (in bp) designated to the left. Lanes 2, 6, and 17 show bands ~ 177 bp, corresponding to vector-vector splicing. Lanes 3, 4, 7, 9, 10, 13, and 18 show bands of various sizes greater than 177 bp. These PCR products contained exons spliced into the pSPL3 splicing vector. Lanes 1, 8, 11, 12, 14, 15, 16, and 19 show no colony PCR product.

Figure 3-2: A Partial Physical Map of the CESCR (redrawn from Johnson et al. 1999)

STSs (or known genes and pseudogenes) are indicated above the thick horizontal double-arrowhead line. A pulsed field gel electrophoresis map is shown below (N = *Not* I A = *Asc* I). The region indicated by A contains Ex 86 (and associated cDNAs); region B contains Ex's 20, 38, and 51; region C contains Ex 60; region D contains Ex's 11A, 41, 45, and 48. Clones in green have been or will be sequenced, those with their addresses in red are near completely sequenced as of August 3rd 1999. Clones starred (☆) were subjected to exon trapping by Dr. Ali Riazi. The CESCR, which extends to between D22S57 and ATP6E, is shown by the arrow located above the map.


Cat eye syndrome critical region
D22S50
D22S601
IGKVP
D22S24
D22S9
D22S789
D22S43
D22S420
D22S137
ATP6E
D22S57
BID
D22S181
Cen
Tel
N A
N
N A
N
N A
N
N
50A16
15J16
87O8
95A8
20K14
120N18
916F2
109L3
213P2
609C6
143I13
238M15
115F6
357F7
233A2
77H2
273A17
1087L10
567H5
200 kb
A
B
C
D

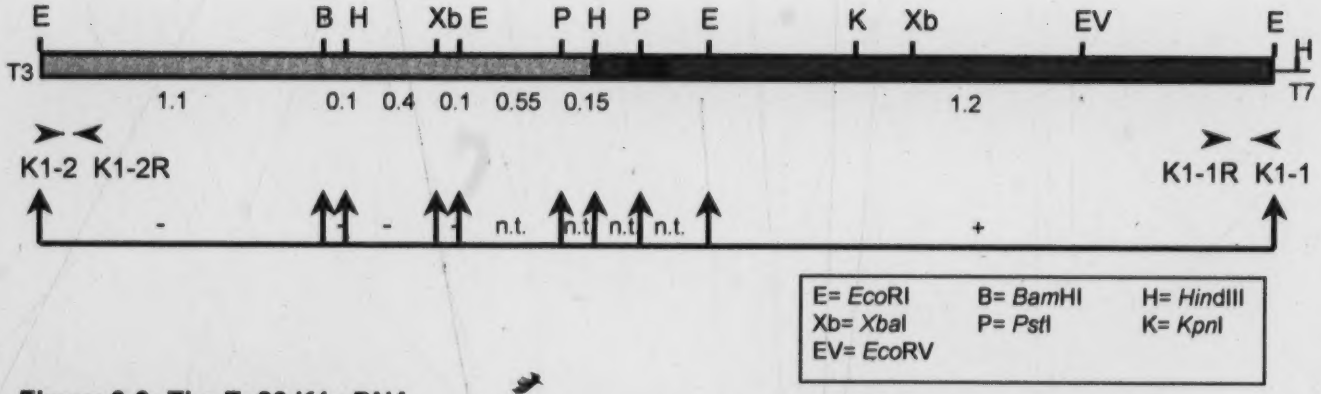


Figure 3-3: The Ex86 K1 cDNA.

This cDNA is approximately 5 kb in size. Restriction enzyme recognition sites and fragment sizes (in kb) are shown above and below the cDNA respectively. The primer pairs used to amplify this cDNA from CESCR PAC or BACs are indicated by the arrowheads. The locations of the T3 and T7 promoter sequences located within the plasmid are shown. A (-) indicates no hybridization of this fragment to CESCR PAC/BACs, a (+) indicates hybridization of this fragment to CESCR PAC/BACs, and (n.t.) indicates that this fragment was not tested as a probe on CESCR PAC/BACs. The vertical arrows correspond with the restriction enzyme sites indicated above the colored line. The location of Ex86 (shown by the blue box) within this cDNA was determined by hybridization of Ex86 to various fragments of the Ex86K1 cDNA.

FIGURE 3-4:Three Ex 86 cDNAs.

Primers used are shown by arrows. K1a, hEST1, and hEST2 were obtained from a fetal brain, parathyroid tumor and combined fetal liver and spleen library respectively. Restriction enzyme recognition sites and/or the approximate size of the fragments are shown above the cDNAs. Ex86-A and Ex86-B are shown by the blue and green boxes respectively. The genomic distance between Ex86-A and the 5' exon (grey box) in hEST1 is 5203 bp. The yellow boxes indicate similarities to various interspersed repetitive DNA elements including a LTR, an Alu, a LINE (L2), and a DNA element (DNA). The K1a cDNA was subcloned into the pBluescript vector using *Hind* III and, hEST1 is contained in the pT7T3D-pac vector and can be excised with *Not* I and *Eco* RI, and hEST2 is in the vector pT7T3D-pac and can be excised with *Pac* I and *Eco* RI.

The two 3' RACE products obtained from hEST1 are shown. 3' RACE product #1 has a 3' splice acceptor site at the same nucleotide position as hEST1 and a 5' splice donor site located 55 bp more 5' than hEST1. 3' RACE product #2 continues past hEST's 5' splice donor for 812 bp, likely running into the intron as a result of priming off contaminating genomic DNA within the RACE kit used.

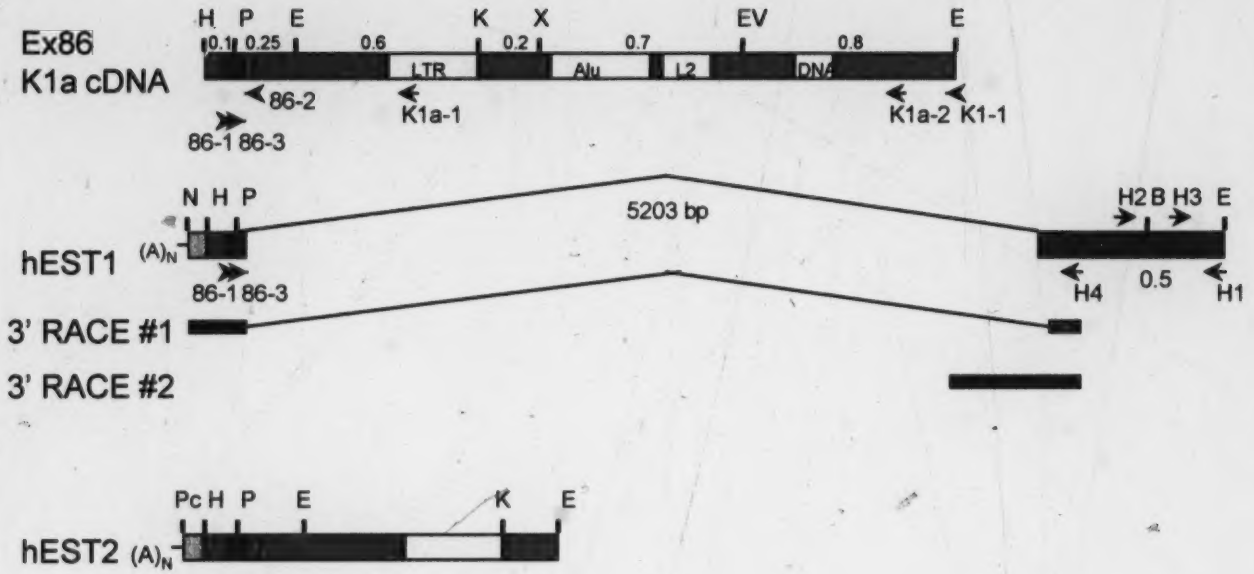

Ex86
K1a cDNA
H P E K X EV E
0.1 0.25 0.6 0.2 0.7 0.8
LTR
Alu
L2
DNA
86-2
K1a-1
K1a-2 K1-1
86-1 86-3
hEST1
N H P
5203 bp
H2 B H3 E
(A)N
86-1 86-3
H4
0.5
H1
3' RACE #1
3' RACE #2
hEST2 (A)N
Pc H P E K E
500 bp
H = Hind III
P = Pst I
E = Eco RI
EV = Eco RV
Pc = Pac I
K = Kpn I
B = Bam HI
X = Xba I
N = Not I
Regions of cDNAs which show identity to each other
Interspersed repetitive DNA
Ex86-A
Ex86-B

Figure 3-5: The sequence of Ex86 K1a cDNA.

The total size of the cDNA is approximately 2600 bp. The sequence of Ex 86 is shown in uppercase letters. The sequences flanking Ex 86 which show both a 3' splice acceptor site and a 5' splice donor site have been underlined. The consensus splice acceptor site sequence is (exon)$G_{100}A_{100}C_{65}$ and the consensus splice donor site sequence is $T_{63}G_{84}A_{68}A_{62}T_{100}G_{100}(G_{73}A_{64}$exon). The restriction enzyme recognition sites have been bolded (AAGCTT=*Hin*dIII, CTGCAG=*Pst*I, GAATTC=*Eco*RI, GGTACC=*Kpn*I, TCTAGA=*Xba*I, GATATC=*Eco*RV). The sequence for the 200 bp region between the *Kpn* I and the *Xba* I sites was not determined and is indicated by a string of 10 N's.

3′

```
   1  aagcttgttg gagaacacca cttatttgga gacaacacgg acttgctgat
  51  gctgccgttg gtgtccttgg cacagGTATA AAGCCCAGCC CTCCGTGCTG
 101  CAGGGACAGG CAGAACACAG ACTTCTGGTG GACCAGCTGG CTCCTAAAGA
 151  GGAGAGGGGA CTAAATCATG TATGCTTGGT TCTTTCTTCC CAGCTTCTCT
 201  CAGgtaactc acgtttcctc ttccccaggt agagagaggc caggagtatt
 251  attaacagaa gtcccatttt cagggaaaca tgactagtag gaaatatgta
 301  tcccttccct acagagaact ccaccttctc agaattctcc ctcagcaaga
 351  ctgagtgtgt atgtgtgtga tgagagagat ggagagagat agaaagagag
 401  agagacagtg agcagtaaac cctcctgcaa gcccttcctg accccagcca
 451  gggggacctg cccatctctc tgtgcctgtt ggccatcctc caccctgtt
 501  cttccaacca cagcttccct gccattcctg tactgggaac caacttctct
 551  tttctttcgt ctgcctccta ctggtgtgag gtaggatatg ccagtttccc
 601  tccttttatc ctttattcct tctcatgtgg aaaagcaacc atgactttta
 651  gccagtcata aggctatatt ttctagcctt ctttgctgct aggcatgacc
 701  atgtggttaa gttttggcca atgaggtgta agcagaagtg gtataagaga
 751  ctccttggaa gtgcacctgt agggaaatgg gataccctt tttgtctctt
 801  tcctggtggc tggaatgtga acataagggc tggagcctga gcagctatct
 851  tagaccatga agtggcagct gtgcagtgaa ggccgtgagg gagcaaaata
 901  ggagcctggg taccNNNNNN NNNNtctaga tgcagacacatttgtgcaag
 951  tttgtggaga aagtaggagt atgaacagtc attcacctgt tgctttttt
1001  ttttttttt tctgagatgg agtctcactc tgtcgcctag gctggagtgc
1051  agtggcacgg tctcggctca ctgcaacctc catctcccgg attcaagcga
1101  ttccccagcc tcggactacc gagtagctgg gattacagat gcgtgccacc
1151  atgcccagct aattttcgga tttttagtag agacgtggtt tcaccttgtt
1201  ggccaggctg gtctcgaact cctgacctca ggtgatctgc cgcctccacc
1251  tcccaaagtg ctgggattac aggccaccgt gctggcctca cctgcctatt
1301  aatgtaaact gagcacccat catgtgcaag gcatcctgca tgacaggact
1351  gtgagggaga caggtgaatc agagagacac tgccctcaag cagcttagag
1401  tccagggcat cgagcttgta cacgtgtgag gccggaactt cccacagtgt
1451  atcttctgtt ctgctacctt tcctgaaaag aactggctca cctcggattc
1501  ctgccaggaa tggctgctgt tttatggggt ttgggccaca actgaggatg
1551  accgaaagca aggcaggcct gggagaaaaa tgtggctcac ttgagatgag
1601  cagggcccac agatatcagt gctggggcac cagttgaaca gaactgtggt
1651  catataggtg tggctggggg ctggggatag agaacaggga gcagcatctg
1701  ttctgggatg ggaagtgtat ggtctccaga acatactggt ccagcgtctg
1751  ctcagagaca tgctaactcc taaaggagag cttcctaaag tgggtcagca
1801  gcatgagcat cctggggagc tgatcagata tgcacatcca aactccaaag
1851  ggagcgccat caatctgggt ttttaggaga cctccaggtg attctgatgg
1901  atgctcaagt tagagaatcc tggacctctg aagccaggta gcaaggtggg
1951  gcaagtcagc gcatagaagg ctacaaattc caaatatagt ataatccaag
2001  caccgaagcc ctctctgaga caaaagctga tttaacaaac cccacgagta
2051  acatgtgatt gaagactggc gaggcccttc taattaacat ccacggtcca
2101  tttgtacaac agacggtaca aaacataatc tggataaact cagaagacgt
2151  gctcatataa agatttggga agtctccaaa ttcaatcaag tgagcagagc
2201  aaatattctg aacctgaata cgaaagaaca tcttaatgtg attgtgtcgt
2251  aacatgactg ggaatcaaaa taaagatgtg ttaagcaaca gcagggcagg
2301  ggttctgctt acaatgtcag ccacagagcc tgagtttgtt agtgtttatt
2351  cagcactgcc caaaggatag agctaggggt ggcagaatat gtgctctgtg
2401  catccggcag ggaattc 5′
```

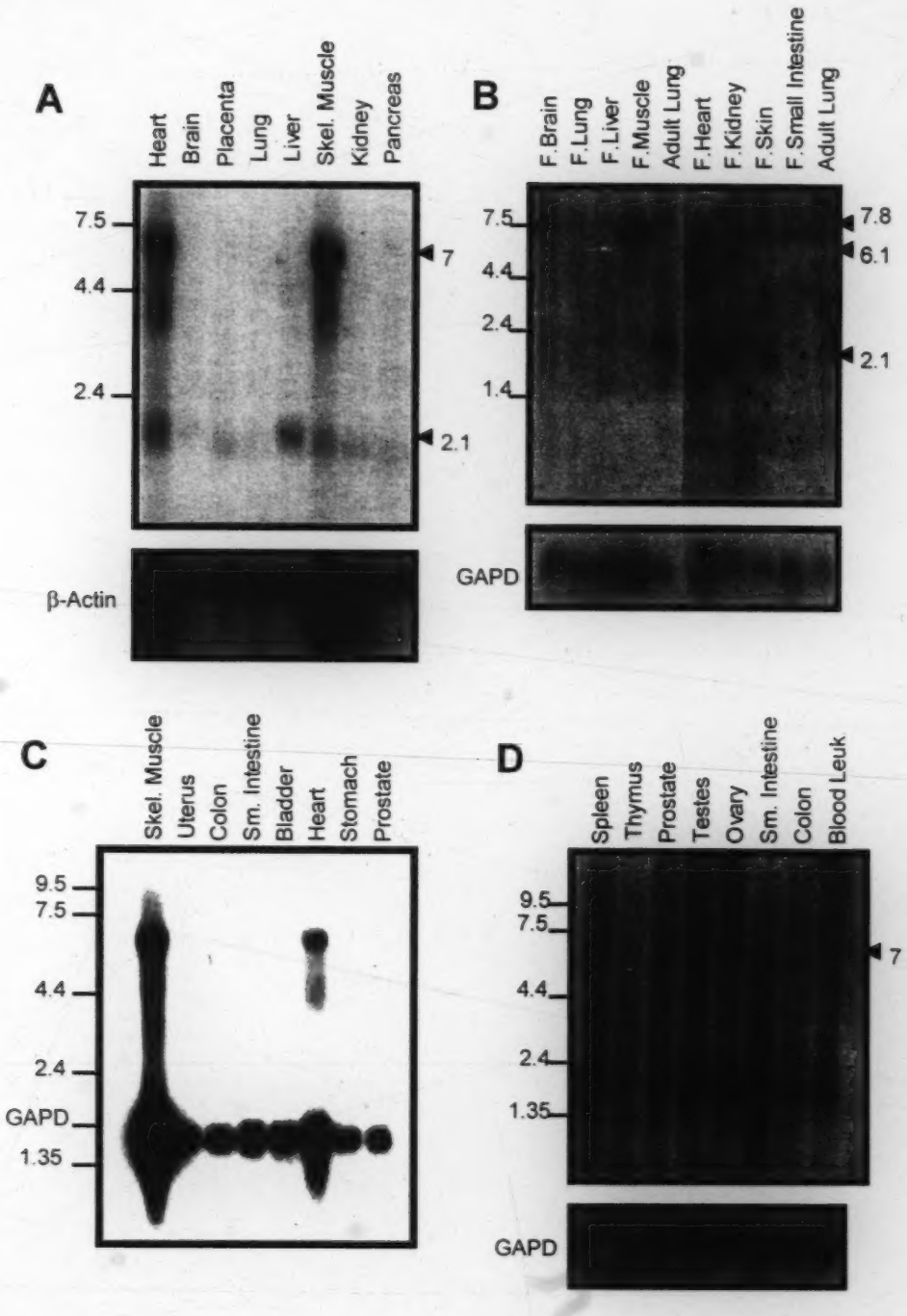
```
    3'
  1 ttttttttttt ttttttttttt ttttttgaga ataaaaatgt cagctttatt
 51 actgctaaaa gcttgttgga gaacaccact tatttgggag caaacacgga
101 cttgctgatg ctgccgttgg tgtccttggc acagGTATAA AGCCCAGCCC
151 TCCGTGCTGC AGGGACAGGC AGAACACAGA CTTCTGGTGG ACCAGCTGGC
201 TCCTtgtact acattcgtgc ctctacctct ccgcagtgcg cgcgggggcg
251 gactcacgtg cgcagtcttc cctatgcgct gggagggcgt cagcaggttt
301 acagaaacaa tgaaagacTT Aacaattggc gcgcgagcag gctcccttga
351 ctggcagggg ccagaggact gaacatcact aagggtctgc cccagcggcc
401 cggccctgaa gtcaaggttg agacccccgg cccagacagt ccgcgaagct
451 gagggaggcg cctggggcct gccgcgcaat gttaccggga tccgcaggca
501 cagggcgata gtacggacgc cgccgcccag gccgtgccca ttgtgcgccg
551 cggccagctc accgcgatct cggtctagga ggcggcggga ccccgcactt
601 cttccttgca acggagcccc tttggcctgc gggtcgagcc tcgtgcc 5'
```

Figure 3- 6: The sequence of the Ex86 hEST1 cDNA.

The total size of the cDNA is 647 bp. The sequence of Ex 86-A is shown in uppercase letters. The consensus polyadenylation signal is underlined at the 3' end. The restriction enzyme recognition sites have been bolded (AAGCTT=*Hin*dIII, CTGCAG=*Pst*I, GGATCC=*Bam*HI). The *Pst*I/*Eco*RI 0.5 fragment used as a probe in Southern and Northern analysis is located from the Pst site in bold print to the 5' end. The ORF of 108 amino acids extends form the 5' end to the capitalized "ATT" stop codon.

Figure 3-7: Expression Analysis of CES86

The probe used for hybridization to these blots was hEST1 *Pst*I/*Eco*RI 0.5. For all panels the tissues are indicated above the pictures ("F." refers to fetal tissues) and sizes (in kb) are designated to the left. The transcript sizes are indicated to the right of each blot. The control probes, either β-actin or GAPD, are shown below the blots (in C, the control is within the picture itself). Panel A shows a Clontech Northern with various adult tissues. A strong transcript ~ 7 kb in size is seen in heart and skeletal muscle. A second transcript ~ 2 kb in size is visible in all tissues. Panel B shows Invitrogen Northern blots with various fetal tissues. A transcript ~ 7.8 kb in size is seen in all tissues except heart, where the transcript is a little smaller, at ~ 6.1 kb. A fainter transcript was also detected at ~ 2.1 kb in most tissues. Panel C shows various muscle-containing adult tissues. A single transcript of 7 kb was seen in skeletal muscle and heart only. Panel D shows various adult tissues. A transcript ~ 7 kb in size was visible in spleen, prostate, ovary, small intestine and colon. A second transcript ~ 2 kb in size was also seen in all tissues.


A
Heart
Brain
Placenta
Lung
Liver
Skel. Muscle
Kidney
Pancreas
7.5
4.4
2.4
7
2.1
β-Actin
B
F.Brain
F.Lung
F.Liver
F.Muscle
Adult Lung
F.Heart
F.Kidney
F.Skin
F.Small Intestine
Adult Lung
7.5
4.4
2.4
1.4
7.8
6.1
2.1
GAPD
C
Skel. Muscle
Uterus
Colon
Sm. Intestine
Bladder
Heart
Stomach
Prostate
9.5
7.5
4.4
2.4
GAPD
1.35
D
Spleen
Thymus
Prostate
Testes
Ovary
Sm. Intestine
Colon
Blood Leuk
9.5
7.5
4.4
2.4
1.35
7
GAPD

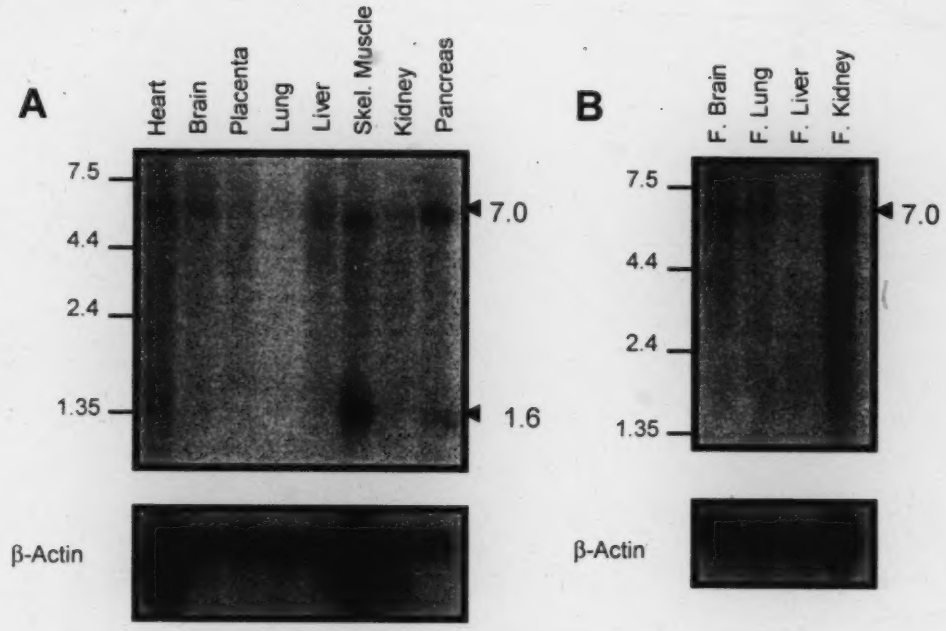


Figure 3-8: Further Expression Analysis of CES86

The probe used for hybridization to these blots was hEST2 PacI/EcoRI 0.4+0.7 (the entire hEST2 insert). The tissues, RNA size ladder (in kb) and control probe are shown at the top, left and bottom, respectively.The transcript sizes are indicated to the right of the blot. Panel A shows a Clontech Northern blot with various adult tissues. A transcript ~7 kb in size is visible in all tissues except lung. A transcript ~ 1.6 kb in size is seen in heart, skeletal muscle and pancreas. Panel B shows a Clontech Northern blot with various fetal tissues. A transcript ~ 7 kb in size was detected in all four fetal tissues.

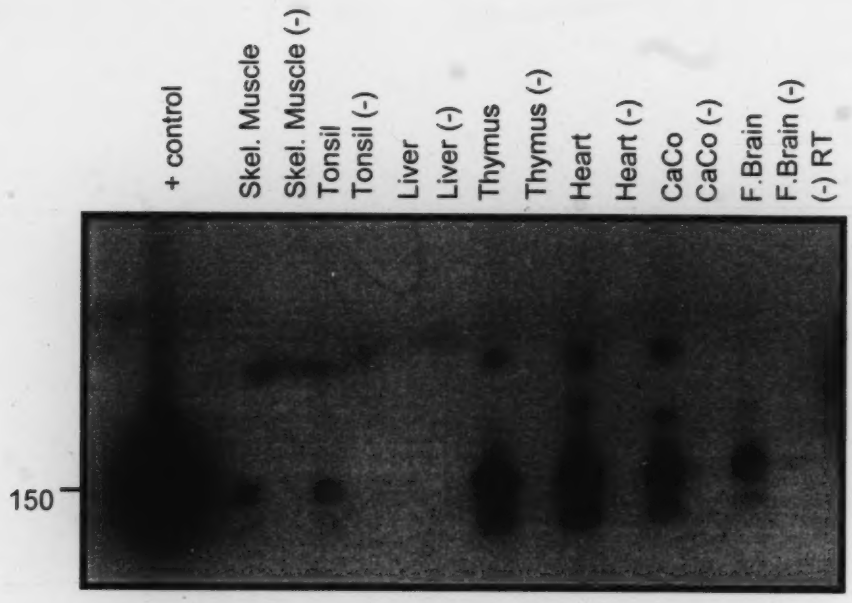


Figure 3-9: RT-PCR Analysis of CES

Total RNA from skeletal muscle, tonsil, liver, thymus, heart, CaCo, and fetal brain was reversed transcribed using the 86-3 primer. Primary PCR was with primers 86-3 and H1 followed by nested PCR with primers 86-3 and H4. The gel was blotted and probed with hEST1 *PstI*/*Eco*RI 0.5. The (+) control consisted of primary and secondary PCR using hEST1 as the template. The (-) controls were identical to the RT reactions but with the omission of Superscript II RT. The (-) RT was identical to the RT reactions but without addition of any template. The expected 150 bp product was detected in all RT reactions except liver. The hybridization of the probe to other bands may be due to alternatively spliced transcripts.

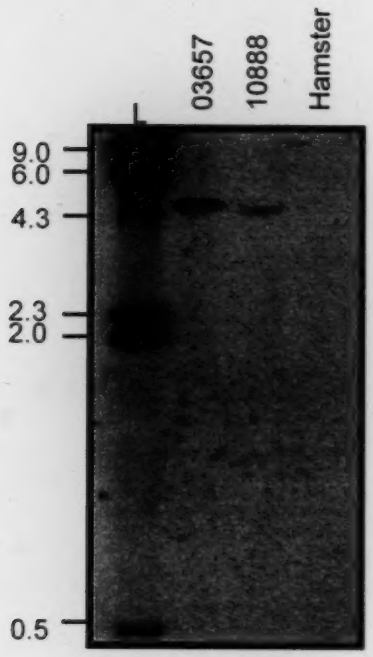


Figure 3-10 : Ex86 Partial Hybrid Panel

Lane L contains a λ*Hin*dIII DNA size marker. Lane 03657 contains total genomic DNA from a normal human cell line. Lane 10888 contains DNA from a human/hamster hybrid cell line containing chromosome 22 as the only human component. The hamster lane contains total genomic DNA from a normal hamster cell line (RJK888). All of these DNAs were digested with *Sst*I. The probe used for hybridization to this Southern blot was Ex86 (isolated from a plasmid). This probe hybridized to a single band in both the total human DNA and the chromosome 22 only DNA.

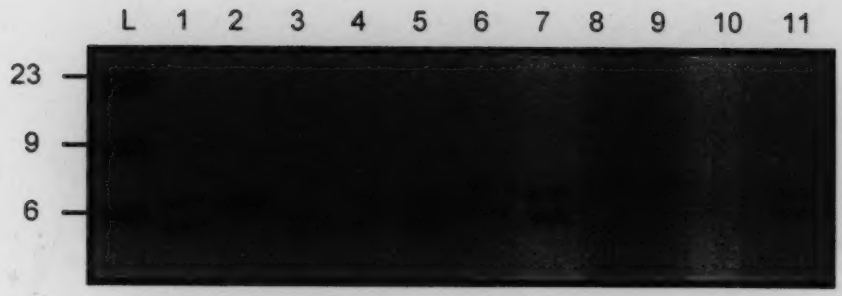
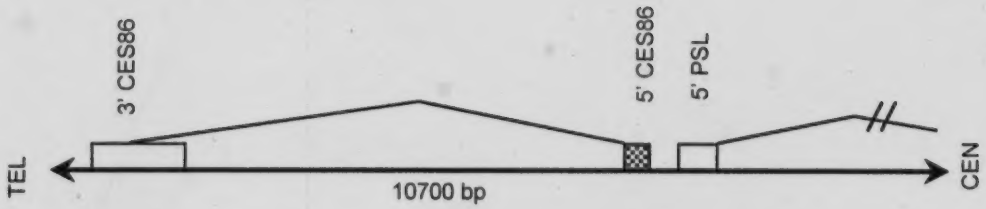


Figure 3-11: CES86 Identifies a Polymorphic *Pst* Site

Lane L contains a λ*Hin*dIII DNA size marker with sizes designated to the left. Lanes 1-11 contain total genomic DNA from 11 different individuals digested with the restriction enzyme *Pst*I. Individuals 2, 3, 8, and 9 (4 out of 11 or 36.4 %) are homozygous, showing only one size fragment, either the larger band or the smaller band. Individuals 1, 4, 5, 6, 7, 10, and 11 (7 out of 11 or 63.4 %) are heterozygous, showing both the larger and the smaller bands together. The allele frequency of the larger allele is 0.6 and the smaller allele is 0.4.

Figure 3-12: Restriction Analysis of CES86 Genomic Region

The *Pst*I/*Eco* RI 0.5 fragment (checkered box) was used to probe several Southern blots containing total human DNA, chromosome 22-only DNA, and total hamster DNA, digested with the restriction enzymes *Bgl* II, *Bsr* GI, *Eco* RV, *Hin* dIII, *Pvu* II, *Ssp* I, and *Sst* I. The expected band size is shown in the chart to the left. The expected bands cover a genomic region of 10700 bp. This region extended into the 3' region of CES86 and proximally into the 5' region of the flanking gene, PSL.



Enzyme	Expected Band size (bp)
Bgl II	5184
Bsr GI	5958
Eco RV	7381
Hin dIII	6886
Pvu II	6058
Ssp I	4000
Sst I	7082

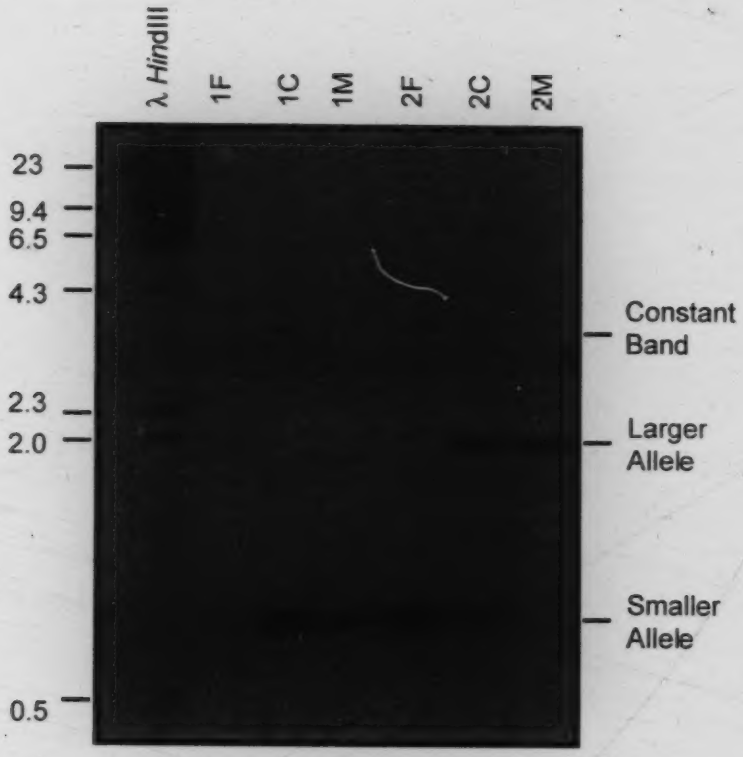
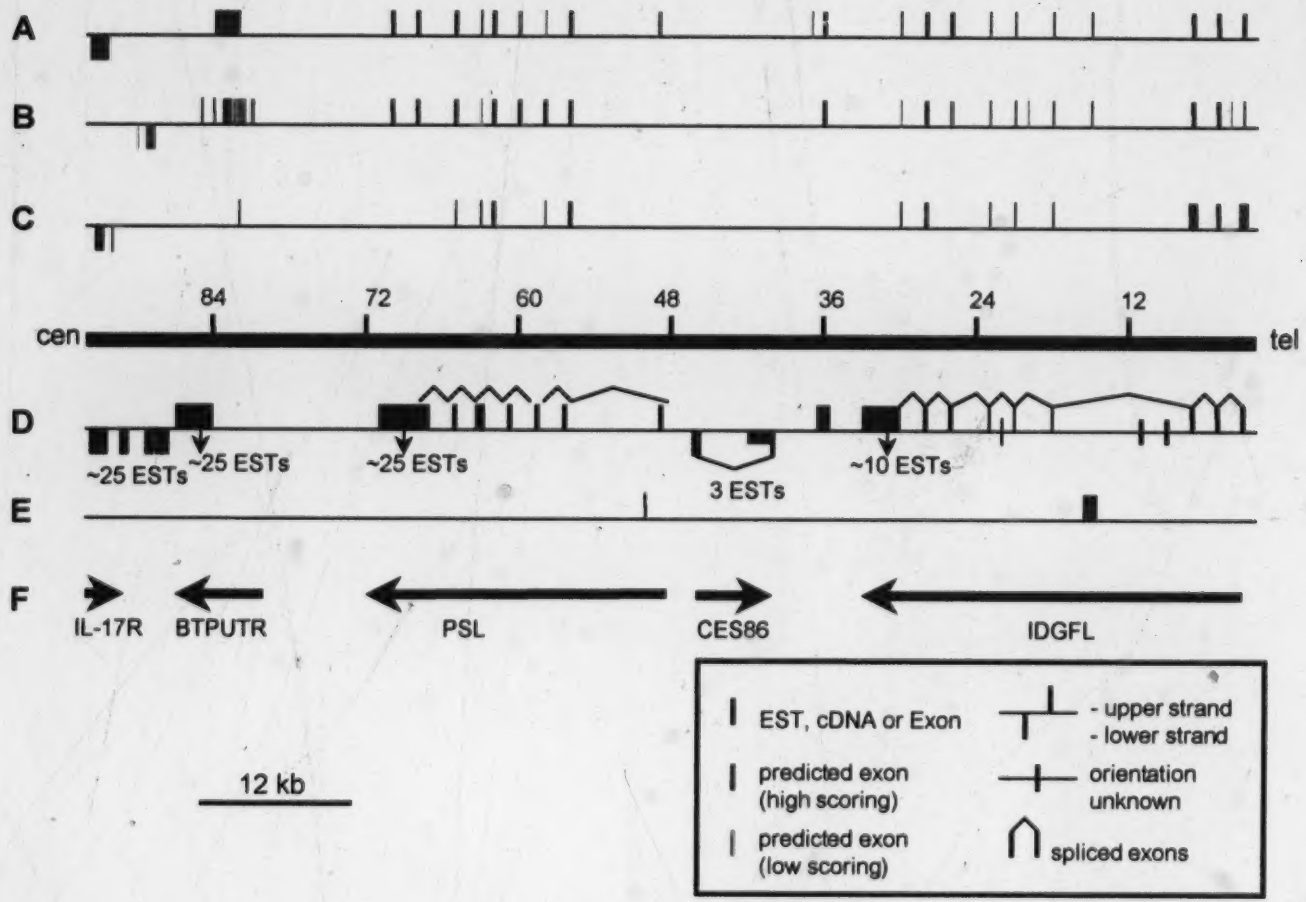


Figure 3-13: Determination of Parent of Origin Using CES86
The sizes (in kb) of the λ*Hin*dIII DNA size ladder are designated to the left. Family one consists of genomic DNA from dad (1F), child (1C), and mom (1M). In this case the child inherits the same allele from both parents and therefore parent of origin of the CES chromosome cannot be determined using this polymorphism. Family two consists of genomic DNA from the father (2F), child (2C), and mom (2M). The child receives one copy of the smaller allele from the dad and three copies of the larger allele from the mom. All DNA was digested with the enzyme *Taq*I.

Figure 3-14: Annotated Sequence of PAC 143i13

Panels A, B, and C represent exons predicted by GENSCAN, GRAIL2, and MZEF, respectively. Panel D represents cDNAs from the EST database, the blue line represents a cloned CpG island. Panel E represents two portions of the sequence which show homology to multiple other sequenced regions of the genome (chromosomes 7, 15, 17, 19, and X). Panel F represents the four putative genes excluding CES86, and their directions (5' → 3'). The thickness of a line representing an exon or cDNA roughly estimates its size. A high scoring exon had a probability ≥ 0.5, while a low scoring exon had a score of < 0.5. Data was compiled from my analysis as well as that of Graham Banting, Tim Footz, Dr. Ali Riazi and Stephanie Maier. All analysis of PAC143I13 was performed between April 8, 1999 and June 16, 1999.


A
B
C
84
72
60
48
36
24
12
cen
tel
D
~25 ESTs
~25 ESTs
~25 ESTs
3 ESTs
~10 ESTs
E
F
IL-17R
BTPUTR
PSL
CES86
IDGFL
12 kb
EST, cDNA or Exon
predicted exon (high scoring)
predicted exon (low scoring)
- upper strand
- lower strand
orientation unknown
spliced exons

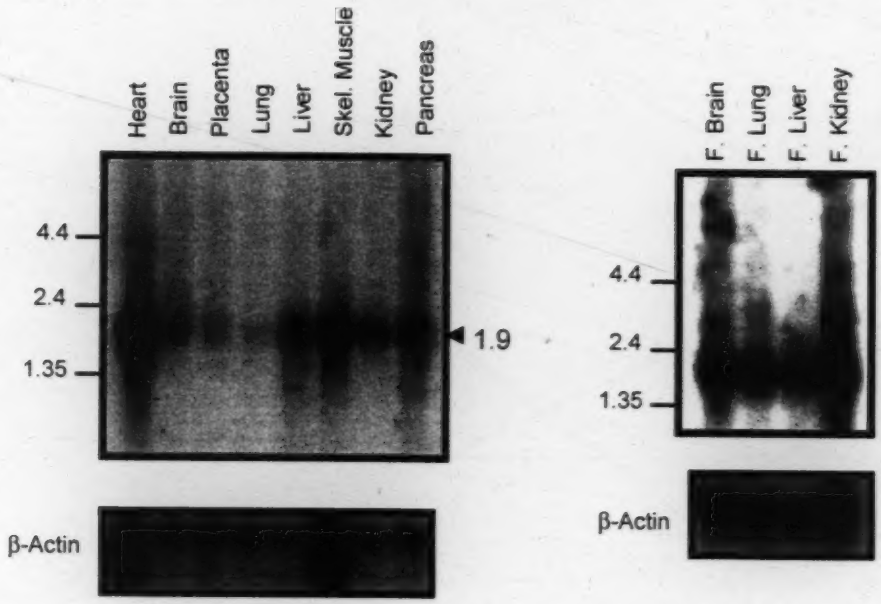
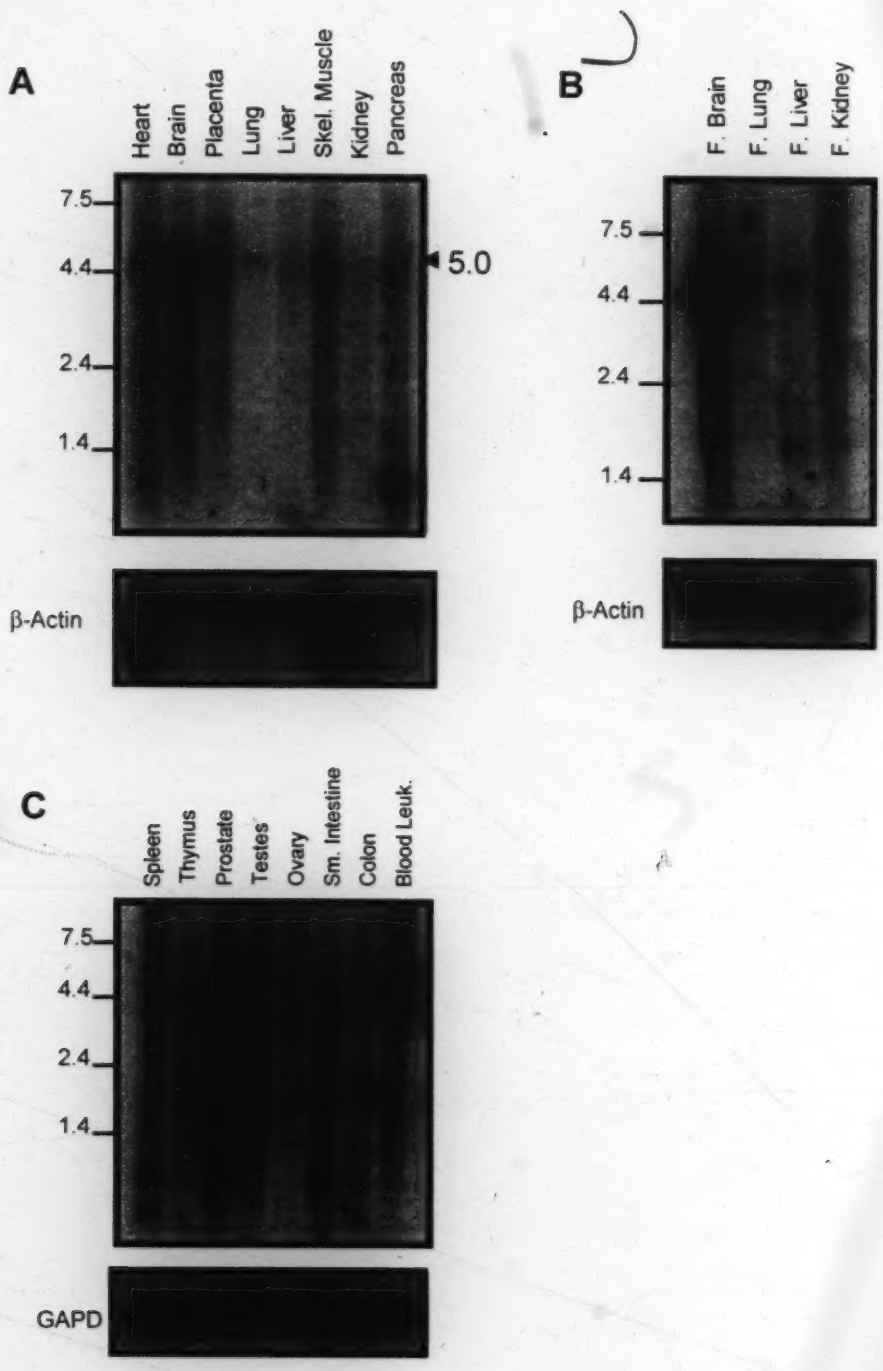


Figure 3-15: Expression Analysis of hEST 52444 (PSL)

The probe used for hybridization to these blots was hEST 52444 (*Pst* 0.8). The tissues, RNA size ladder (in kb), and the control probe are shown at the top, side and bottom, respectively. Panels A and B show Clontech Northern blots with various adult or fetal tissues. An abundant transcript of approximately 1.9 kb in size can be seen in all adult and fetal tissues.

Figure 3-16: Expression Analysis of hEST 46414 (BTPUTR)

The probe used for hybridization to these blots was hEST 46414 *Hin*dIII/*Not* 1.5. Panels A and C show various adult tissues and Panel C shows fetal tissues. A transcript ~ 5 kb in size was strongly expressed in heart, brain, prostate, testes, peripheral blood leukocytes, and fetal brain and weakly expressed in all other tissues.


A
Heart
Brain
Placenta
Lung
Liver
Skel. Muscle
Kidney
Pancreas
7.5
4.4
2.4
1.4
5.0
β-Actin
B
F. Brain
F. Lung
F. Liver
F. Kidney
7.5
4.4
2.4
1.4
β-Actin
C
Spleen
Thymus
Prostate
Testes
Ovary
Sm. Intestine
Colon
Blood Leuk.
7.5
4.4
2.4
1.4
GAPD

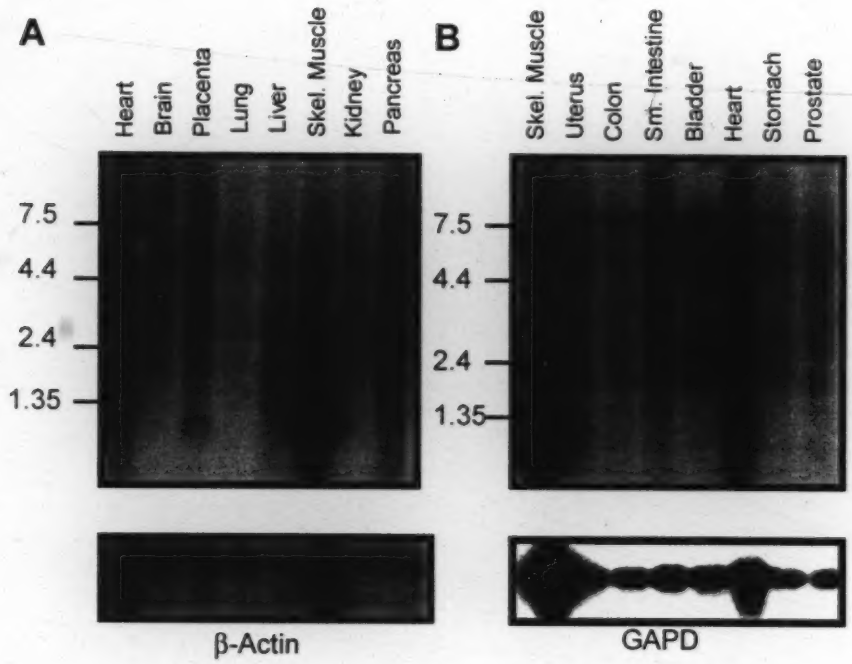
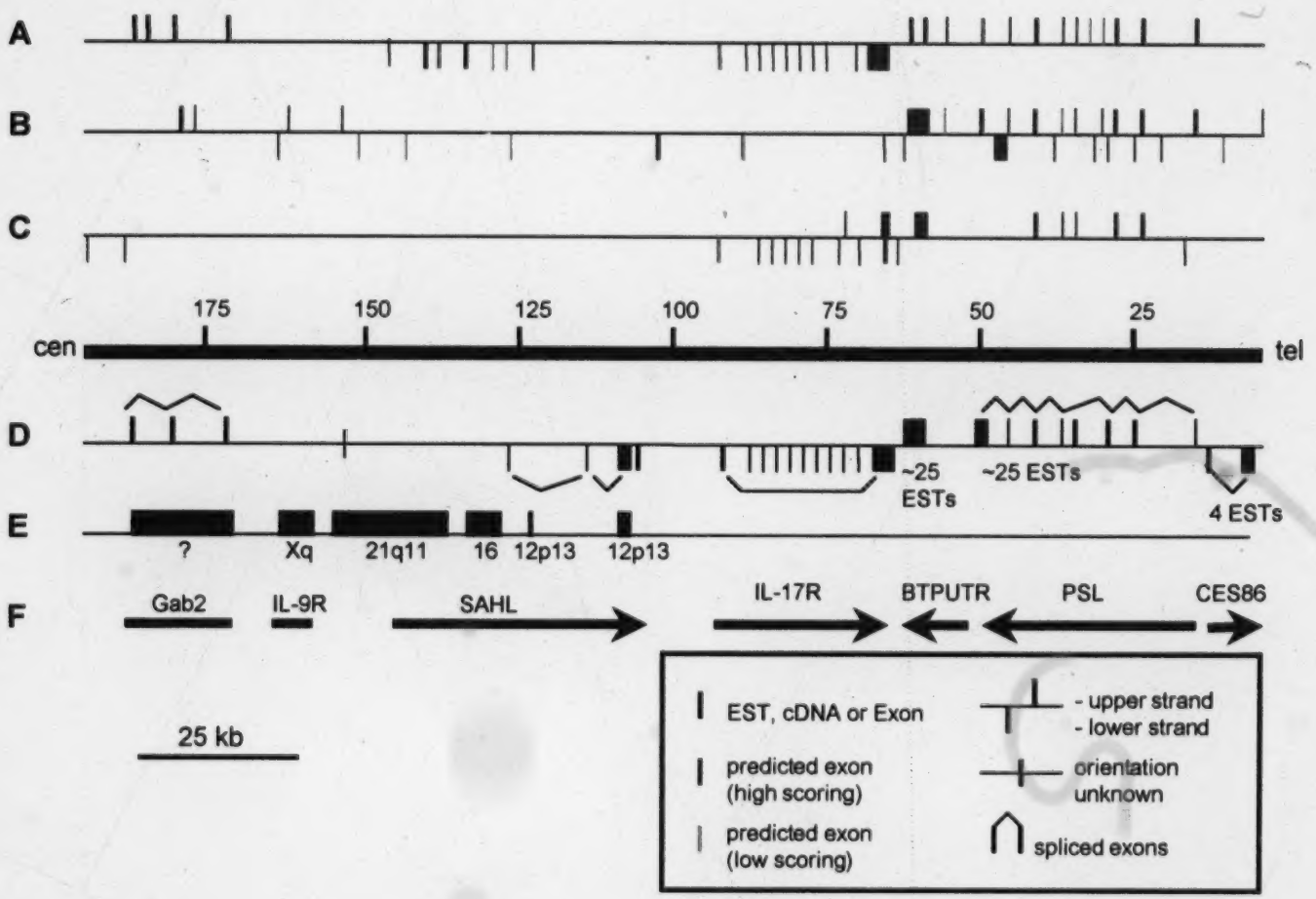


Figure 3-17 :Expression Analysis of IL-17 Receptor Gene

The probe used for hybridization was a 697 bp PCR product. Panels A and B show mRNA from various adult tissues. Various sized transcripts (8.8, 6.3, 5.0, 2.6, 1.9, and 1.05) can be seen in all lanes. Autoradiographs provided by Stephanie Maier, used with permission.

Figure 3-18: Annotated Sequence of PAC 109L3

Panels A, B, and C represent exons predicted by GENSCAN, GRAIL2, and MZEF, respectively. Panel D represents cDNAs from the EST database. Panel E represents portions of the sequence which show similarity to several other sequenced regions of the genome (indicated below the line). Panel F represents the putative genes (CES86, PSL, BTPUTR, SAHL and the published sequence of IL-17R) and pseudogenes (Gab2 and IL-9R)and their directions. All analysis of PAC 109L3 was performed on August 31, 1999.


A
B
C
175
150
125
100
75
50
25
cen
tel
D
~25 ESTs
~25 ESTs
4 ESTs
E
?
Xq
21q11
16
12p13
12p13
F
Gab2
IL-9R
SAHL
IL-17R
BTPUTR
PSL
CES86
25 kb
EST, cDNA or Exon
predicted exon (high scoring)
predicted exon (low scoring)
- upper strand
- lower strand
orientation unknown
spliced exons

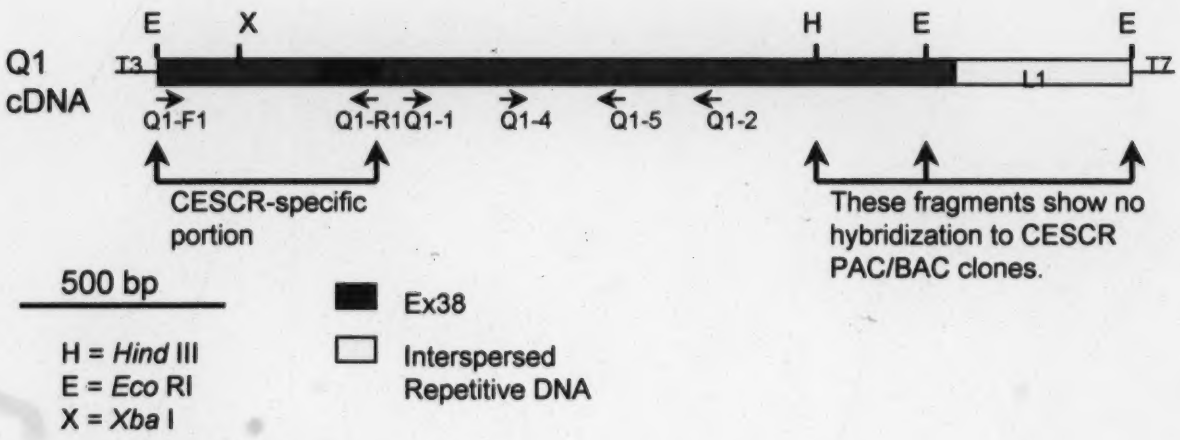


Figure 3-19:The Ex 38 cDNA.

Primers used are shown by arrows. This cDNA was obtained from a fetal brain cDNA library. The cDNA is contained within the pBluescript vector and can be excised by *Eco* RI. The complete sequence of this cDNA can be found in Figure 3-21. The primers Q1-F1 and Q1-R1 were used to generate the PCR product used as a probe for Southern and Northern analysis.

Figure 3-20: The Sequence of the Ex38 Q1 cDNA

The total size of the cDNA is 2208 bp. The portion of the cDNA which maps to PAC 238M15 is bolded. The sequence of Ex 38 is in uppercase letters. The ORF of 55 amino acids extends form the 5' end to the underlined "TGA" stop codon. The F1/R1 PCR probe used for Southern and Northern analysis is bolded.

5′

```
   1  agtattctac agatgtctgt taggtctagc agtttatagt actgttcaag
  51  tcctctattc cttgttgatc ttctatctaa ttgttctgtc cactgtatac
 101  aatgaagtat ttaagctctc ataattattg ttgggttgtc tatttttact
 151  tcaattctgt cagttactgc ttcataaagt ttgaggctct gttattaggt
 201  gcctatatgt ttatacgtgt tatgtctcct cgttggattt gacttttttt
 251  attgttttaa aagtttcttt gtgcctagat actcgatttt ttgaatctgt
 301  cgatgatgtt gtcattggaa attttatcct aatcccgaag tactgttcac
 351  taacctcatc attaacaact tactatattg tttttgaacc aatcatttaa
 401  acggtgtgac acaacacgat taacgttccg ttgtgaattc atacccccag
 451  gaagaatgac taaatctgtt ttaagtgtct ttgtcctctt tctttactga
 501  ttgtgtcagg tgacctttct tagccttcca aaccagctgg ctgaagttat
 551  tttaggctaa cttgttttcc tgaaatagta ctgtttgttc aaaataaagt
 601  atttgagaaa gtcttttaca acatgaaaga cctcaagcag attcaaatat
 651  aacccatctc cctctcattc aataattaat tttaggaaaa agcttgccta
 701  atattcaagc gcgcgcagtc tccagctgta tttgctgttt gttttattct
 751  gcttcacata gaggttcagc tgatcacaca ccatgctgta tctactcctt
 801  ctcctctggt tttctggtgc tttgaaagtc acctccgacc cctgaggcac
 851  tggagcttcc cagtcgaaag tcctggtgag tagcatgtgg gttccagtga
 901  gacatgcagt tgatgtctaa gcctcaggag gatgtcagct gggcctcctg
 951  atcaacactc cctccaaacc ttgtcagata actaacaggc ttggtcagtt
1001  aagcaggagg cttgcaaacc aacagaagca tagccctgtg ttccataaag
1051  attaattacc ttcaggacac caaagaggtg tgggtcattc aaggacagaa
1101  tcaccattct tggccccatt gctccttggc tctgggccac tgggtctgaa
1151  ccctttgaaa ggcattttca gagccatgtg tttaggagaa aagcccatct
1201  gtacactcaa actctctctc acacacacac acacactcac actcacatag
1251  tggttggtgt tcttcatgga gtcctagaat ctcaaggctg gcagggccaa
1301  gacattctct cagtacataa ccaatggctc cttctgtttc attcacatgg
1351  aatatgtggc tagggccaaa gatacaggct tatttttcaa tgtgtgtatc
1401  caaacgtgac tgagaaatgc acattttgaa aaaccagacc atctggcttt
1451  aaaatacata ttgaaagtaa gagtgccagt gaggttgtca agtacacact
1501  gtgttcagga cttctgattt accgactcct gcttgtcctg attctgcatc
1551  tctcctgctt tcaatcatca ttctccctct ttagaagagc aacaggcaga
1601  aatgaacatg ttttgcctga attaaaataa caaaaccaaa accgaaaact
1651  agactcaaag agacttcgta cctgtgaatt aGAGCCTGGG AGTGAGACAG
1701  AAGCAGCTGG GAGGAGAAAA ATCAACGCGC CAGGGGCCCT ACTGGTCCTT
1751  CTGCCTTAGC CACAGGTTCT GAAACCAAAG CAAAACCACC AGAGAGTGAT
1801  TCATGTGGgt atggcctcc ctccccctgc ccagaaaggt ttctcttgag
1851  cccatatgct agtttctgtc atgcaaacgc ccctggcaat gccgtccttg
1901  ccttggcgca atggctctga tggggtaggc atctccgggg gttaatactg
1951  gcacaaactc atgtaaatct tcctgggcct aagtctggtg tactctcaat
2001  gtgaattggc atggtcctgg agggctagtc agctagctcc tagcctcgag
2051  atcggcctgg accaaaccaa aaaaaaaaat catctctggt gaaacaggat
2101  agccacccca ccccctgctc agcattcccc agtgcctccc ttcacagtgg
2151  ccagtttgga ttcttctcac caaacgctca agacctgtac aggttactta
2201  ttggtcc 3′
```

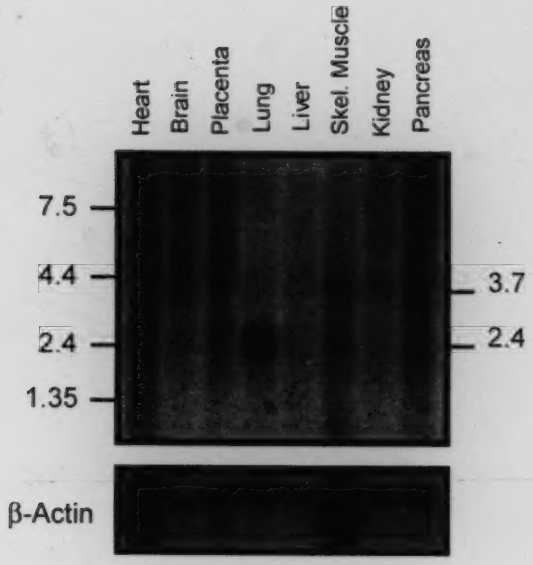


Figure 3-21 : Expression Analysis of CES38

The probe used for hybridization to these blots was the CES38 F1/R1 PCR product. The above blot contained mRNA from various adult tissues. A transcript ~ 2.4 kb in size was strongly expressed in lung while a fainter transcript ~ 3.7 kb in size was seen in all tissues tested. This result was obtained subsequent to thesis experiments.

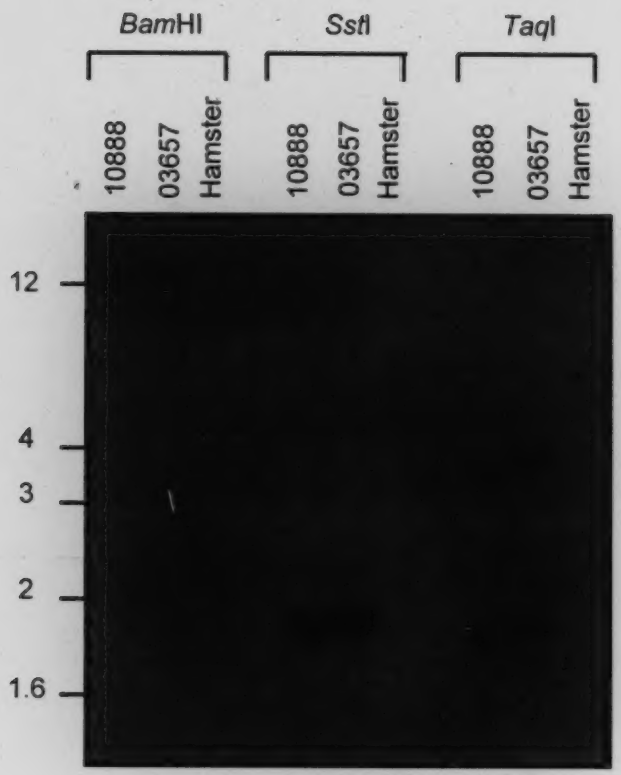
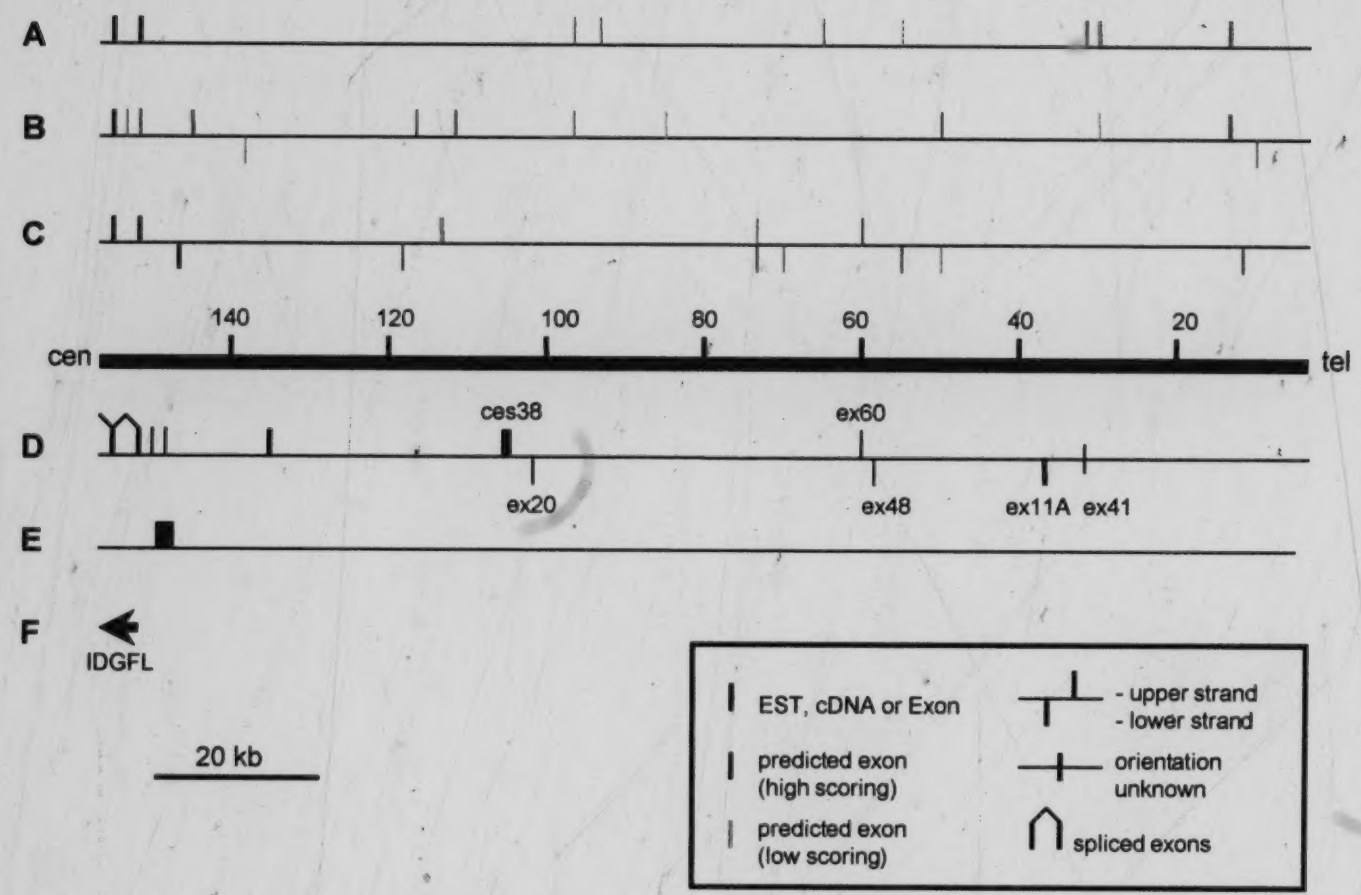


Figure 3-22: CES38 Partial Hybrid Panel

Lanes 10888 represent DNA isolated from a human/hamster hybrid cell line containing human chromosome 22 as its only human component. Lanes 03657 represent DNA isolated from a normal human cell line containing total genomic DNA. Hamster lanes represent a normal hamster cell line containing total genomic DNA. These DNAs were digested with *Bam*HI, *Sst*I, and *Taq*I (from left to right on the picture). The probe used for hybridization to this blot was the CES38 F1/R1 PCR product. Locations of DNA size markers are designated to the left.

Figure 3-23: Annotated Sequence of PAC 238M15

Panels A, B, and C represent exons predicted by GENSCAN, GRAIL2 and MZEF, respectively. Panel D represents cDNAs from the EST database and exons 38, 20, 60, 11A, 41, and 48 obtained from exon trapping. Panel E represents a portion of the sequence which shows homology to several other sequenced regions of the genome (Chromosomes 1, 10 and 16). Panel F represents the one known gene partially on this PAC and its direction (5' → 3'). The thickness of a line representing an exon or cDNA roughly estimates its size. All analysis of PAC 238M15 was performed on May 24, 1999.


A
B
C
140
120
100
80
60
40
20
cen
tel
D
ces38
ex60
ex20
ex48
ex11A
ex41
E
F
IDGFL
20 kb
EST, cDNA or Exon
predicted exon (high scoring)
predicted exon (low scoring)
- upper strand
- lower strand
orientation unknown
spliced exons

GENOMIC CLONE	# pAMP10 CLONES	# 177 bp PRODUCTS	# NO PCR PRODUCTS	# PUTATIVE EXONS	# NOVEL EXONS
BAC 95A8	27	15	10	2	0
PAC 109L3	13	5	7	1	1
PAC 238M15	79	16	30	33	8
totals	119	36	47	36	9
PAC 120N18	Clones from this PAC were not characterized				
BAC 233A2	Clones from this PAC were characterized by Dr. Ali Riazi				

Table 3-1: Summary of Exon Trapping Results

The total number of pAMP10 clones analyzed was 119. Of these, 36 were 177 bp products produced from vector-vector splicing, and 47 yielded no colony PCR product. A total of 36 putative exons were found and once sequenced, nine novel exons were isolated.

EXON	SEQUENCE	BLAST RESULT
11A	CTTCATCAACTTCATTCCTGAGTCATACTTACT TAATAAACAAGACAAGATCAACAACAAGAAA GGTGGAAGAACACAGCTGGTTTACCTGCACA GGAGCTAAATACTTTGCAATTCCATTGGCTGA GCGCAACACCAAGAGGCTGACTAAGAGGAG CACACATGCACAACTGCTGCGTGGGAAACAG GATGGCAGCGA (201 bp)	♦ No similarities to previously known genes or ESTs ♦ No repetitive elements
20	GGAGTCACTGACGGTTGGAGGACTGATATTC ACCAATACCTTTCCCACAAACAAGAGATGCCT GAGAAAGACCTGGTCATGCAGGAGGTG (93 bp)	♦ No similarities to previously known genes or ESTs ♦ No repetitive elements
38	GAGCCTGAGAGAGAGACAGAAGCAGCTGGG AGGAGAAAAATCAACGCGGCCAGGGGCCCT ACTGGTCCTTCTGCCTTAGCCACAGGTTCTGA AACCAAAGCAAAACCACCAGAGAGTGATTCA TGTGG (128bp)	♦ No similarities to previously known genes or ESTs ♦ No repetitive elements
41	GGGCCTGCAGAGGGAGGCAGCCGCTGTTGG GAGGTCACAGAGCGTCTGCAGCTGACTGGG AAGCCCTCCTAACTGGCGCACCTGAGGGGCT GGGTGCCGTCTGCTGCTTCTGGCTGCCCTGG CCCGCAGTAGT (133 bp)	♦ No similarities to previously known genes or ESTs ♦ No repetitive elements
45	GCTGGTCTCAAACTCCTGGGCTCAAGCAATTC TCCAGCCTCAAACTGTGCTGGGATTACAG (61 bp)	♦ No similarities to previously known genes or ESTs ♦ 61 of 61 bp are an Alu repetitive element
48	GGTCAGGAATCCAGGCATGGCTTAGGTGGC CTCTCCTTAGGTTGCAGTCAAGATGTGAGCTA GAGTTTTAGTCCAGTCTGGGAAGTGAAATCC CATCCATTTCGCCCCTCTCGTCAGTGAGGCGT TTGAGTCCAGAAAGATGAG(145bp)	♦ No similarities to previously known genes or ESTs ♦ 81 of 145 bp are part of a LTR element
51	GGCTTCTCATCCTCAGCACTGCTGACATTTGC GCCAGATATTTCTTTGTTTTGGAGGCTGTCGT CTGCACTGTCGGATGTTTAGCAGCATCGTCA GCCTCTACTCTCCA (110bp)	♦ No similarities to previously known genes or ESTs ♦ 108 of 110bp are part of a MER1 DNA element
60	ATCAACACACAAAGACAATTGATCTCGAACCA GCTTCCTACACTATCTTG (50 bp)	♦ 100% identical to a previously trapped exon (Trofatter, 1995). ♦ No repetitive elements
86	GTATAAAGCCCAGCCCTCCGTGCTGCAGGGA CAGGCAGAACACAGACTTCTGGTGGACCAGC TGGCTCCTAAAGAGGAGAGGGGACTAAATCA TGTATGCTTGGTTCTTTCTTCCCAGCTTCTCTC AG (127 bp)	♦ Similarity to 3 ESTs: W35386 (hEST1), AA704966 (hEST2), AA505576 ♦ No repetitive elements

Table 3-2: A List of the Putative Exon Sequences and an Interpretation of the BLASTN and TBLASTX Search Results.

EXON	ENZYMES TESTED	# LOCI
11A	*Sst, Pst*	1
20	*Sst*	1
38	*Sst*	1
41	*Sst*	1
45	*Sst, Pst*	Smear
48	*Sst, Pst*	Possibly 2
51	*Sst*	Smear
60	*Sst*	No Result
86	*Sst*	1

Table 3-3: A List of the Putative Exons and the Number of Loci they Likely Represent in the Genome.

The Southern blots used contained DNA from a normal human cell line (GM03657), a chromosome 22-only cell line (GM010888) and normal hamster DNA along with molecular size markers. Some Southern blots contained these DNAs digested with the restriction enzymes *Sst* I and *Pst* I while some contained these DNAs digested with the restriction enzyme *Sst* I only. Exons which likely represent one locus in the genome showed the same number and size of band(s) in the total human DNA lane as in the chromosome 22-only DNA lane. Exon 48 hybridized to two bands in the total human DNA, only one of which hybridized in chromosome 22-only DNA, raising the possibility that it may represent more than one locus. Exons 45 and 51 showed smears on the autoradiograph, likely because they are both entirely composed of repetitive DNA. No result was obtained using Exon 60, probably because it is small (50 bp) and hybridized poorly.

CHAPTER 4: DISCUSSION

A) Identification of Genes in the CESCR

The main purpose of this study was to identify genes located in the minimal critical region for CES. The region of focus was an approximately 1 Mb region located about 1 Mb from the centromere. This region was defined by comparing the duplication overlap of two CES patients, one carrying a supernumerary r(22) chromosome and one harbouring an interstitial duplication (Mears et al. 1995). One plausible explanation for the etiology of the abnormal features seen in individuals with CES is the overexpression of a gene or genes located within the CES duplication region. Such a gene(s) would function in the development of the affected organs/tissues such as heart, eyes, urogenital system, outer ears, face and brain.

Studies in gene-rich regions of the human genome have suggested a gene density of one per 10-50 kb (Ansari-Lari et al. 1997) and one per 20-25 kb (Gong et al. 1996). Using an average of one gene per 40 kb, then the 1 Mb distal CESCR is expected to contain about 25 genes. Since the CESCR is located within the gene-poor pericentromeric region of 22q, the number of expected genes may be somewhat lower.

The only gene to date reported in the CESCR is ATP6E (Baud et al. 1994). This gene localizes to the most distal portion of the CESCR and codes for the E subunit of the vacuolar H^+-ATPase proton pump, which appears to be ubiquitously expressed. Although the effect of overexpression of ATP6E on the features of CES cannot be excluded, the widespread expression of this housekeeping gene is unlikely to be

responsible for the CES phenotype. Therefore this study was done to isolate and characterize more candidate genes from the CESCR.

1) Amplification of Putative Gene fragments

A number of cosmid, PAC, and BAC genomic clones were subjected to the exon trapping procedure by Dr. Ali Riazi. A portion of these isolated exons were further analyzed in this study by sequencing, hybridization to cDNA libraries, Southern and Northern blots, and computer sequence analysis and database searching. Overall the exon trapping procedure was successful in identifying several gene fragments (exons) and two putative genes, CES38 and CES86.

- **Exon Trapping Summary**

Table 3-1 provides a summary of the exon trapping results. A total of 119 putative exon-containing pAMP10 clones were analyzed; 27 from BAC 95A8, 13 form PAC 109L3, and 79 from PAC 238M15. Of these 119 clones, 36 or 30.3 % were due to vector-vector splicing and produced a 177 bp product when colony PCR was performed. Vector-vector splicing occurs when the exon trapping vector splices the portions of the HIV *tat* exons together without an insert. The exon trapping vector is designed such that when this occurs a *Bst*XI site is created at the junction of the two exons. Such products can be cleaved with the restriction enzyme *Bst*XI producing two very small fragments which will not be subcloned into the pAMP10 cloning vector. Since 30.3 % of the pAMP10 clones tested by PCR showed the 177 bp product produced by vector-vector splicing, it is obvious that the enzymatic treatment with *Bst*XI was not fully effective at

eliminating such clones. This could have been due to a short incubation time or perhaps the enzyme was old and had lost some of its effectiveness. Alternatively, the *Bsf*XI restriction site at the junction of the two HIV exons may have been aberrant due to PCR artifacts. This source of contamination could have been reduced by isolating the putative exon-containing secondary PCRs from a low-gelling temperature agarose gel and then subcloning these into pAMP10 (i.e.: rather than subcloning the entire secondary PCR reaction, which contained many vector-vector splicing products).

Another source of contamination were clones which gave no result when colony PCR was performed. This was seen for 47 out of 119 clones, corresponding to 39.5 % of all clones analyzed. This result is theoretically inconsistent with the exon trapping procedure. Such contamination may have been due to satellite colonies that were not ampicillin resistant and did not contain a plasmid. Small satellite colonies are often observed when bacteria are grown on ampicillin containing plates in which the ampicillin is at a low concentration or has degraded over time. This source of contamination could be reduced by careful plating of bacteria on freshly made LB+ampicillin plates. It is also possible that due to the age of the colonies examined, some of them had lost their plasmids. Therefore no PCR product could be amplified.

- **Effectiveness of Exon Trapping**

The partial or complete genomic sequence for many of the PAC/BAC clones in the CESCR is now available and has allowed identification of genes by computer prediction programs. Figure 4-1 shows a comparison of computer predicted genes within PACs 109L3, 143I13, and 238M15 and exons identified by exon trapping from PACs

109L3 and 238M15 (only those examined in this study). Since no novel exons were discovered from BAC 95A8, the genomic sequence in this region was not annotated. IL-17R is a multiexon gene present on 109L3 which was predicted well but from which no exons were trapped. BTPUTR is a putative single exon gene which was predicted well but which would not be discovered by the exon trapping procedure. PSL is a multiexon gene which was predicted well but from which no exons were trapped. CES86 is a part of a putative gene with two identified exons, one of which was partially trapped due to the presence of cryptic splicing signals on the opposite DNA strand. CES38 is a portion of a putative gene with one trapped exon and an associated cDNA, which was not identified by gene or exon prediction. IDGFL is a multiexon gene which was predicted well, and from which Dr. Ali Riazi characterized one partial exon, Exon B (Riazi, 1998; Riazi et al. submitted). Therefore two out of the five, or 40 %, of the predicted genes were confirmed by trapped exons. The six exons trapped from 238M15 suggest, by their directions, that there may be up to four genes on this PAC, none of which were computer predicted. It may also be possible that any of these exons were trapped due to the presence of cryptic splicing signals and therefore would not likely represent transcribed DNA sequences. However, exons 11A, 20, 41, and 60 each contain an ORF in at least one frame, suggesting that they may represent real exons.

Genomic sequence analysis is another gene identification technique being used in the CESCR. This technique is much easier and less time consuming than many other traditional techniques, but genomic sequence may not be available for a region for some time after the region has been cloned. In the meantime traditional techniques must be used. Genomic sequence analysis seems to be an effective method to identify most genes in a region. The predicted exons/genes must be further

confirmed by cDNAs, either from ESTs, RT-PCR or RACE. Disadvantages of this technique include inaccurate prediction of exons, lack of prediction of some exons, and prediction of exons within pseudogenes. One of the putative genes examined in this study, CES38, has not been predicted by any gene/exon prediction program used thus far. As well, putative 5' exons of CES86 were only predicted by GRAIL2. This supports the idea that multiple gene identification techniques must be used to identify all genes in a given region. Using any one gene identification technique will usually identify only a subset of all genes in a region, as is seen here.

- **Other Possible gene Identification Techniques**

Other gene identification techniques which could have been used in this region include CpG island cloning, cDNA selection and comparative mapping (currently underway by T. Footz). CpG island cloning is perhaps a technically less demanding approach but is only applicable to genes that have CpG islands. Analysis of genomic sequence using the computer program GeneTool allows identification of putative CpG islands.

No CpG islands were predicted in or near IDGFL, CES38 or trapped exons present on PAC 238M15. Therefore it is highly unlikely that these genes would have been identified by the CpG island cloning technique. In contrast, two CpG islands were predicted near the 5' end of PSL, suggesting that this putative housekeeping gene may have been identified by CpG island cloning. A third CpG island was predicted within the body of the PSL gene. It is unlikely that this island is associated with PSL since housekeeping genes usually have their CpG islands at the 5' end. It is possible that this island may be associated with one of the surrounding genes, possibly CES86, if this gene is in fact an active gene. A CpG island was also predicted at the 5' end of the BTPUTR gene,

indicating that this gene may also have been identified by the CpG island cloning technique. Likewise for IL-17R, a CpG island was predicted, covering the initial exon.

Another gene identification technique that could have been used is cDNA selection. This technique is expression-dependent and therefore it may be difficult to identify cDNAs that are tissue or time specifically expressed or at very low abundance. The putative genes IL-17R, BTPUTR, PSL, and IDGFL likely would have been identified by cDNA selection since they each have a fairly widespread expression pattern, and each identifies numerous ESTs. In contrast, CES38 shows strong expression only in lung and therefore may not have been identified by cDNA selection unless lung transcripts were examined. Likewise, CES86 is strongly expressed only in heart and skeletal muscle. Since we are looking for developmentally important genes which are active from the third to eighth weeks of fetal development, cDNA selection may not have identified genes which may be expressed only at this time.

- **Drawbacks of Exon Trapping Within the CESCR**

During the time that Dr. Riazi was performing exon trapping on various genomic clones from the CESCR, we first became aware of the large number of unprocessed, truncated pseudogenes in the 22q pericentromere. As well, recent sequence analysis of a number of genomic clones in the proximal half of the distal CESCR has revealed the presence of multiple duplicated gene fragments from chromosomes 2, 10, 11, 12, 16, X, and Y (McDermid et al. 1999 and unpublished results). Several duplicated gene fragments are now known to reside within the pericentromeric region of 22q and therefore within the CESCR. These include fragments of the NF1 gene (Regnier et al. 1997), the vWF gene

(Eikenboom et al. 1991), the GGT family of genes (Courtay et al. 1994), the IGKV3 gene (Lotscher et al. 1986), the ALD gene (Eichler et al. 1997) the KCNMB3L gene (Riazi et al. submitted) and many other gene fragments (Minoshima et al., 1998; McDermid, unpublished results). Exons and cDNAs corresponding to these gene fragments may be easily identified by exon trapping (Riazi, unpublished results). It is for this reason that exon trapping should be cautiously used as a method of gene discovery in pericentromeric regions.

2) Identification of a Putative Gene, CES86

CES86 was identified by exon trapping Ex 86 and subsequently obtaining a cDNA from a fetal brain library and three ESTs from the EST database. Northern hybridization with one of the ESTs detected an abundant transcript of approximately 7 kb in heart and skeletal muscle. Southern hybridization and PCR place CES86 in the central portion of the distal CESCR near locus D22S43. Preliminary Southern hybridization with Ex 86 suggested that it represented a single-copy locus in the genome. Initial Southern hybridization with the *Pst*I/*Eco*RI fragment from hEST1 suggested that it represented more than one locus in the genome, but further analysis refuted this. It was therefore hypothesized that CES86 represented a duplicated, unprocessed 5' truncated pseudogene. Comparison of all the experiments presented in this thesis along with genomic sequence analysis of PAC 109L3 now suggest most strongly that CES86 is in fact an active and single copy gene within the CESCR. The possibility of CES86 being a truncated and unprocessed, duplicated

gene fragment has not been entirely eliminated, therefore a discussion of several possibilities will be covered.

Hypothesis #1: CES86 is an Active Gene in the CESCR

Several experiments and observations have suggested that the partial CES86 gene identified to date is in fact part of an active, single copy gene located within the CESCR. The CES86 gene is represented by four ESTs which match identically the genomic sequence. Northern hybridizations performed with a portion of one of these ESTs (hEST1) identified a 7 kb transcript strongly expressed in heart and skeletal muscle. RT-PCR analysis of hEST1 confirmed its expression in several tissues. As well, 3' RACE experiments detected this transcript in adult brain. Southern hybridizations support this hypothesis, as total human DNA and chromosome 22 only DNA digested with seven different restriction enzymes (*Bgl* II, *Bsr* GI, *Eco* RV, *Hin* dIII, *Pvu* II, *Ssp* I, and *Sst* I) show the identical number and size of bands. The same DNAs digested with *Pst* I and *Taq* I showed two bands in the total human lane, only one of which was seen in the chromosome 22 only lane. These extra bands were shown to be polymorphic restriction sites. Genomic sequence analysis of PAC 109L3 has identified several predicted exons within the PSL genomic region but on the opposite DNA strand, which could reasonably be a part of the CES86 gene. GRAIL2 predicted seven high scoring ($p>0.5$) exons and one moderate scoring ($p<0.5$) exon, while MZEF predicted one high scoring exon. As well, the predicted CpG island within the body of the PSL gene may be associated with CES86.

Several observations provide evidence against the hypothesis that CES86 is an active gene in the CESCR. Southern hybridization performed by T. Footz shows no evidence of a mouse ortholog. He has found that the

one known gene in the CESCR, ATP6E, plus at least eight other putative genes (BID, GAB, MTP, CTCO, CES38, PSL, BTPUTR, and IL-17R) map to mouse chromosome 6. Thus, nine out of the ten genes identified so far have apparent mouse orthologs. This region has been cloned into BACs and a number are currently being sequenced. The mouse genomic sequence around CES86 is not yet complete, suggesting that once complete sequence is available there may be regions of orthology to CES86.

As well, it is puzzling why no further 5' cDNA sequence has been obtained by either ESTs, cDNA library screening or RACE. This may be explained by the high CG content in the area surrounding the most 5' exon. This area was predicted to be a CpG island for PSL. The high CG content may easily hinder the effectiveness of reverse transcriptase enzymes, leading to prematurely truncated cDNAs, RT-PCR or RACE products.

After analysis of the genomic sequence of PAC 143i13, it seemed as if there was no apparent room for the CES86 gene. At the most 5' end of CES86 there are only 52 bp until a cDNA-confirmed exon of PSL. This suggested that if CES86 were a real gene then its 5' end could extend off PAC 143I13 and into PAC 109L3. However, this meant that it must contain at least three genes within one or more of its introns, one of which, IL-17R, is in the same orientation and transcribed from the same DNA strand as CES86. Analysis of the recently available complete sequence of PAC 109L3 shows no prediction of exons centromeric of IL-17R which could reasonably be a part of the CES86 gene. However, several exons were predicted just proximal to CES86 and overlapping with the PSL gene, but on the opposite DNA strand.

Hypothesis #2: CES86 is a 5' Truncated Gene Fragment Duplicated From Another Chromosome

This hypothesis was put forth because of the lack of further 5' cDNA sequence. If this hypothesis were correct it must be confirmed by probing a complete somatic cell hybrid panel. This experiment was attempted six different times on two differently digested hybrid panels, with no satisfactory results. Further evidence supporting this hypothesis would be some degree of sequence divergence (depending on the age of the duplication) between the observed cDNAs, which would be transcribed from the active gene copy, and the chromosome 22 genomic sequence. Such divergence of sequence was not observed, suggesting that the CES86 cDNAs are transcribed from this region.

Hypothesis #3: CES86 is a 5' Truncated Gene Fragment Duplicated From Elsewhere on Chromosome 22

This hypothesis is highly unlikely but was put forth because of the lack of further 5' cDNA sequence and the lack of alternatively sized (or number of) bands on a genomic Southern. If the CES86 gene were a very recently duplicated gene fragment then little or no sequence divergence would be expected, thus conserving all restriction fragment lengths. If this hypothesis were correct it could not be confirmed by probing a complete somatic cell hybrid panel. Alternatively, a specific chromosome 22 hybrid panel would be necessary to map the active copy. The active copy could also be located by similarities to sequenced genomic clones elsewhere on chromosome 22. Considering that most of this chromosome is now sequenced, completion expected before the end of 1999, a second locus would most likely have already been identified.

Future Experiments

Several experiments could be repeated to confirm the results that were obtained in this study and to potentially clone the remaining portion of the CES86 gene. The suggested experiments will be discussed in the context of the three hypotheses presented.

Hypothesis #1: CES86 is an Active Gene in the CESCR

The 5' RACE experiment from hEST1 should be repeated, perhaps using different primers. As well, the PCR could be manipulated to optimize the conditions for CG-rich regions. In order to confirm that the Ex86 K1a is a real mRNA and not genomic contamination, RT-PCR must be done with mRNA that has been treated with DNase in order to eliminate the possible amplification from genomic DNA.

RT-PCR could be performed between the GRAIL2 exons predicted proximal to CES86 and between these predicted exons and the 3' end of CES86. RT-PCR conditions should be optimized to allow the reverse transcriptase enzyme to continue through the CG-rich region of CES86. If further 5' cDNA sequence can be obtained, it should be used to probe Northern blots to confirm the expression data presented here. Once the structure of CES86 has been determined, it should be subjected to protein structure analysis to identify any conserved domains and to detect similarities to previously identified genes. This will help us determine whether this gene could be a candidate for some of the phenotypes associated with CES. The effects of overexpression of this gene could be further studied by creating BAC transgenic mice and observing for any abnormal phenotype.

Hypothesis #2: CES86 is a 5' Truncated Gene Fragment Duplicated From Another Chromosome

To test this hypothesis, CES86 must be re-probed onto a complete monochromosomal hybrid panel. As well, as the sequence of the human genome is completed, the CES86 probe could be tested by submitting the sequence to BLAST searches of the htgs, monthly and nr databases. This would identify any genomic clones and their chromosomal locations showing similarity to CES86.

Hypothesis #3: CES86 is a 5' Truncated Gene Fragment Duplicated From Elsewhere on Chromosome 22

To test this hypothesis several experiments could be performed. The PAC 143i13 could be used as a probe for fluorescence in-situ hybridization (FISH). This experiment would only be informative if the region which was duplicated was significantly distal to the CESCR, which could be resolved by FISH analysis. The CES86 gene could be used as a probe on a chromosome 22 hybrid panel. This hybrid panel consists of various hybrids that contain only portions of chromosome 22 (Budarf et al. 1996), such that the probe would map to the CESCR and some other region. As well, continued genomic sequencing of chromosome 22 should allow identification of the region which was duplicated to the CESCR. This would be identified by continued BLAST searches of the htgs, monthly and nr databases using CES86.

3) Identification of a Putative Gene, CES38

CES38 was identified by exon trapping Ex 38 and subsequently obtaining a cDNA from a fetal brain cDNA library. Southern hybridization placed CES38 in the central portion of the distal CESCR near locus D22S43 but distal to CES86. Characterization of the Q1 cDNA suggested that it was chimeric to an unknown extent. Further characterization of this putative gene was continued when the partial genomic sequence of PAC 238M15 was available and subsequent to other thesis experiments. The 22-specific portion was amplified by PCR and used as a probe on a partial hybrid panel. This experiment suggested that the 22-specific portion represented only one locus in the genome. Northern analysis with this PCR probe detected an abundant 2.4 kb transcript in adult lung and a fainter 3.7 kb transcript in several other tissues. It is puzzling why no Northern hybridizations produced successful results using the Q1 cDNA. It is possible however, that the large chimeric portion of this cDNA somehow hindered the 528 bp, 22-specific portion from adequately hybridizing. As well, Northern hybridization on Clontech multiple tissue Northern blots was attempted only once. The other Northern hybridizations with the Q1 cDNA were performed on laboratory made Northern blots containing mRNA from two cancer cell lines only.

Future Experiments

Further characterization of this putative gene should continue by first cloning the remainder of the gene. This could be accomplished by performing 5' and 3' RACE or screening more cDNA libraries. As well, since one of the trapped exons (Exon 60) contained within PAC 238M15 is in the same orientation as CES38 it would be worthwhile performing

RT-PCR between these regions, perhaps using mRNA from adult lung, where CES38 was most highly expressed. The genomic sequence analysis (using GENSCAN, GRAIL2, and MZEF) which has been performed to date on PAC 238M15 has predicted only low to moderate scoring exons. Predicted gene elements with such poor probabilities are likely to be inaccurate, if real at all. Therefore more sequence analysis with different exon prediction programs could be performed before RT-PCR between such predicted exons is performed.

In such cases where gene structure cannot be easily elucidated from the human genomic sequence, it may be advantageous to make use of the mouse genomic sequence. Analysis of partial mouse genomic sequence has revealed similarities between the human CES38 and the putative mouse Ces38. The regions of similarity are both 5' and 3' of the human Ex38Q1 cDNA and show similarities between 73% and 93%. RT-PCR between these conserved regions in human is currently underway. T. Footz has isolated by PCR a mouse region showing similarity to human CES38. This PCR product should be used to study the expression pattern in mouse tissues and the results compared with the CES38 human Northern obtained. As well, computer gene prediction can be performed using mouse genomic sequence. Subsequently RT-PCR between predicted mouse exons can be performed.

Once the structure of CES38 has been determined, it should be subjected to protein structure analysis to identify any conserved domains and to detect similarities to other known proteins. This will help us decide whether this gene is a good candidate for CES. Transgenic mice have already been produced harbouring the human PAC 238M15. These mice, so far, appear normal but should be examined for the expression of the human CES38 mRNA and protein, especially in the lungs.

4) Partial Characterization of a Putative Gene, BTPUTR

BTPUTR was identified by genomic sequence analysis of PAC 143I13. Analysis of the ESTs for BTPUTR and the computer predicted exons as well as ORF analysis has lead to the hypothesis that this is a putative single exon gene. This gene shows two ORFs, one of which shows similarity to transcriptional factor motifs. Northern hybridization showed strong expression of a 5 kb transcript in prostate and brain and weaker expression in many other tissues.

The theory that this putative gene is a single-exon gene must be confirmed by RT-PCR and/or RACE. Northern hybridization with the 5' coding region of this gene should be performed and compared with the Northern results previously obtained. Identification of the 5' end by RT-PCR or RACE may help in elucidating which of the ORFs is used, or if both are used. However, protein studies through Western analysis may be necessary to determine which ORF is used. The gene structure may also be examined by comparing the mouse genomic sequence to the human genomic sequence to see which regions are conserved. Since one of the putative ORFs codes for a transcription factor, it is a good candidate for CES. It should therefore be tested to see if this mRNA and protein are overexpressed in CES individuals. The effects of overexpression of this gene could be further studied by creating BAC transgenic mice and observing for any abnormal phenotype. If this putative gene does encode a transcription factor, it would be interesting to find out what gene(s) it regulates.

5) Partial Characterization of a Putative gene, PSL

PSL was first identified by genomic sequence analysis of PAC 143I13. By comparing the ESTs for PSL with the computer predicted exons, it appears to be composed of at least 8 exons. Analysis of the ORF shows similarity to proteins from several organisms including yeast phosphatidyl synthase. Northern hybridization detected an abundant 1.9 kb transcript in all tissues tested. Although the overexpression of PSL on the features of CES cannot be excluded, the widespread expression of this probable housekeeping gene is unlikely to be responsible for the CES phenotype. Therefore, past confirming the predicted gene structure, few experiments will likely be performed.

Developmental Defects in CES, a Molecular Etiology

As previously stated, the organs most often affected in CES include the eyes, outer ears and face, heart, kidney, anus and brain. Understanding the molecular etiology of this syndrome therefore requires a detailed knowledge of the affected structures. The development of the heart, eyes, ears, face, and urogenital system all occur between weeks three to eight of embryonic development. Therefore it is likely that CES is a result of abnormal gene expression during the first few critical weeks of development. Since CES is associated with a duplication of a portion of chromosome 22, the most plausible explanation for the phenotypic effects is the overexpression of a dosage sensitive gene or genes located within the duplicated region. Such genes may be directly or indirectly involved with the development of the affected organs. The products of such dosage sensitive genes may belong to a group of inherently dosage sensitive

functions like ligand/receptor or signal transduction molecules, transcriptional regulation, structural proteins and morphogens.

A possible but less likely hypothesis explaining the phenotype associated with CES involves the high number of duplicated gene fragments located within the 22q pericentromeric region. Exons from various gene fragments, in the presence of promoter-type sequences, may be transcribed into a novel gene product. Overexpression of such a gene could theoretically cause the developmental defects associated with CES. If such a fusion-type gene were the cause of CES, it would make studies of model organisms, such as the mouse, difficult as such a gene would likely have arisen in the human genome after the evolution of man and mouse.

Future Research

Identification of Genes From the CESCR

Since complete genomic sequence of the CESCR, in both human and mouse, is expected by the end of 1999, most gene identification will be most easily performed through computer analysis of genomic sequence and comparison of the two genomes. This eliminates the need for further exon trapping or other traditional gene discovery techniques such as cDNA selection or CpG island cloning.

Further Characterization of CES38 and Other Candidate Genes

Further characterization of candidate genes may provide insight into the molecular etiology of CES. The first step in characterization is to isolate full length cDNAs for such genes. This may be accomplished by

cDNA library screening, EST searches, RT-PCR and RACE. Subsequently, the expression profiles of all genes must be determined by hybridization to Northern blots containing mRNA from various adult and fetal tissues. With the increasing availability of mouse genomic sequence, mouse orthologs not already identified can be discovered and more detailed expression can be studied in the mouse using Northern analysis and *in-situ* hybridizations. Access to genomic sequence will also allow identification of regulatory regions including promoters. Characterization of the putative proteins can be performed by computer analysis of sequence. The cDNA sequence can be used to search for important structural motifs such as DNA binding domains often found in transcription factors, or hydrophobic membrane-spanning regions often found in membrane receptors. The presence of such motifs may give clues as to the function of the protein. As well, the overexpression of these genes in CES patients must be confirmed by finding overexpression of the mRNA by Northern analysis and overexpression of the protein by Western analysis.

CONCLUSIONS

By using exon trapping and genomic sequence analysis as methods of gene discovery within the minimal duplicated region causing CES, two putative genes and several gene fragments were discovered. In order to determine whether these putative genes are possible candidates for some of the features of CES, they must be further characterized as described.

The results presented in this thesis stress the need for more than one gene identification technique to be employed for any given region. The CES86 putative gene was partially identified by both exon trapping and genomic sequence analysis while the putative CES38 gene was identified by exon trapping but not by analysis of genomic sequence. As well, CES38 is not represented by any ESTs, highlighting the fact that not all genes will be represented by ESTs.

Figure 4-1: Exon Trapping Versus Genomic Sequence Analysis in a Portion of the CESCR

PACs 109L3 and 238M15 were subjected to the exon trapping procedure and thereby identified 7 exons (86, 38, 20, 60, 48, 11A, and 41; ExB was characterized by Dr. Ali Riazi) as shown. The exons obtained from BAC 95A8 are shown since they were identical to two exons already being characterized by Dr. Ali Riazi. The genomic sequences of PACs 109L3, 143i13, and 238M15 were analyzed for the presence of computer-predicted exons and genes. This method identified five putative genes (IL-17R, BTPUTR, PSL, CES86 and IDGFL) as shown. Only CES86 and IDGFL were identified by both gene identification techniques. Up to four genes may be represented by the six novel exons located on PAC 238M15. The direction (5'→3') of the gene or exon is indicated by the arrow representing that gene or exon. The direction of Ex41 is unknown as it is not flanked by the conserved 3' and 5' splicing signals.

Cen
109L3
143i13
238m15
Tel
Genes identified by Genomic Sequence Analysis
IL-17R
BTPUTR
PSL
CES86
IDGFL
Genes identified by Exon Trapping
Ex 86
Ex B
CES38
Ex 20
Ex 60
Ex 48
Ex 11A
Ex 41

REFERENCES

Aalfs, C.M., Fantes, J.A., Wenniger-Prick, L.J., Sluijter, S., Hennekam, R.C., van Heyningen, V., and Hoovers, J.M. (1997) Tandem duplication of 11p12-p13 in a child with borderline development delay and eye abnormalities: dose effect of the PAX6 gene product? *Am J Med Genet* **73(3)**:267-71.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215(3)**:403-10.

Ansari-Lari, M.A., Shen, Y., Muzny, D.M., Lee, W., and Gibbs, R.A. (1997) Large-scale sequencing in human chromosome 12p13: experimental and computational gene structure determination. *Genome Res* **7(3)**:268-80.

Aurias, A., Rimbaut, C., Buffe, D., Zucker, J.M., Mazabraud, A. (1984) Translocation involving chromosome 22 in Ewing's sarcoma. A cytogenetic study of four fresh tumors. *Cancer Genet Cytogenet* **12(1)**:21-5.

Barry, A.E., Howman, E.V., Cancilla, M.R., Saffery, R., and Choo, K.H. (1999) Sequence analysis of an 80 kb human neocentromere. *Hum Mol Genet* **8(2)**:217-27.

Baud, V., Mears, A.J., Lamour, V., Scamps, C., Duncan, A.M., McDermid, H.E., and Lipinski, M. (1994) The E subunit of vacuolar H(+)-ATPase localizes close to the centromere on human chromosome 22. *Hum Mol Genet* **3(2)**:335-9.

Bell, C.J., Budarf, M.L., Nieuwenhuijsen, B.W., Barnoski, B.L., Buetow, K.H., Campbell, K., Colbert, A.M., Collins, J., Daly, M., Desjardins, P.R., DeZwaan, T., Eckman, B., Foote, S., Hart, K., Hiester, K., Van Het Hoog, M.J., Hopper, E., Kaufman, A., McDermid, H.E., Overton, G.C., Reeve, M.P., Searles, D.B., Stein, L., Valmiki, V.H., Watson, E., Williams, S., Winston, R., Nussbaum, R.L., Lander, E.S., Fishbeck, K.H., Emanuel, B.S., and Hudson, T.J. (1995) Integration of physical, breakpoint and genetic maps of chromosome 22. Localization of 587 yeast artificial chromosomes with 238 mapped markers. *Hum Mol Genet* **4(1)**:59-69.

Berger, R., Bernheim, A., Weh, H.J., Flandrin, G., Daniel, M.T., Brouet, J.C., Colbert, N. (1979) A new translocation in Burkitt's tumor cells. *Hum Genet* **53(1):**111-2.

Bernardi, G. (1995) The human genome: organization and evolutionary history. *Annu Rev Genet* **29**:445-76.

Berry, A. C. (1987) Rubinstein-Taybi syndrome. *J. Med. Genet.* **24**: 562-566.

Bird, A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature* **321(6067):**209-13.

Bird, A.P., Taggart, M.H., Nicholls, R.D., and Higgs, D.R. (1987) Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. *EMBO J* **6(4)**:999-1004.

Brown, W., and Tyler-Smith, C. (1995) Centromere activation. *Trends Genet* **11(9)**:337-9.

Buckler, A.J., Chang, D.D., Graw, S.L., Brook, J.D., Haber, D.A., Sharp, P.A., and Housman, D.E. (1991) Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. *Proc Natl Acad Sci U S A* **88(9)**:4005-9.

Budarf, M.L., Eckman, B., Michaud, D., McDonald, T., Gavigan, S., Buetow, K.H., Tatsumura, Y., Liu, Z., Hilliard, C., Driscoll, D., Goldmuntz, E., Meese, E., Zwarthoff, E.C., Williams, S., McDermid, H., Dumanski, J.P., Biegel, J., Bell, C.J., and Emanuel, B.S. (1996) Regional localization of over 300 loci on human chromosome 22 using a somatic cell hybrid mapping panel. *Genomics* **35(2)**:275-88.

Buetow, K.H., Weber, J.L., Ludwigsen, S., Scherpbier-Heddema, T., Duyk, G.M., Sheffield, V.C., Wang, Z., and Murray, J.C. (1994) Integrated human genome-wide maps constructed using the CEPH reference panel. *Nat Genet* **6(4)**:391-3.

Buhler, E.M., Mehes, K., Muller, H., and Stalder, G.R. (1972) Cat-eye syndrome, a partial trisomy 22. *Humangenetik* **15(2):**150-62.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78-94.

Burge, C.B., and Karlin, S. (1998) Finding the genes in genomic DNA. *Curr Opin Struct Biol* **8(3)**:346-54.

Burn, J., Takao, A., Wilson, D., Cross, I., Momma, K., Wadey, R., Scambler, P., and Goodship, J. (1993) Conotruncal anomaly face syndrome is associated with a deletion within chromosome 22q11. *J Med Genet* **30(10)**:822-4.

Burset, M., and Guigo, R. (1996) Evaluation of gene structure prediction programs. *Genomics* **34(3)**:353-67.

Carlock, L., Wisniewski, D., Lorincz, M., Pandrangi, A., and Vo, T. (1992) An estimate of the number of genes in the Huntington disease gene region and the identification of 13 transcripts in the 4p16.3 segment. *Genomics* **13(4)**:1108-18.

Chance, P.F., Alderson, M.K., Leppig, K.A., Lensch, M.W., Matsunami, N., Smith, B., Swanson, P.D., Odelberg, S.J., Disteche, M., and Bird, T.D. (1993) DNA deletion associated with hereditary neuropathy with liability to pressure palsies. *Cell* **72(1)**:143-51.

Choo, K.H.A. (1997) Centromere DNA dynamics: latent centromeres and neocentromere function. *Am J Hum Genet* **61**:1225-1233.

Choo, K.H.A. (1998) Turning on the centromere *Nat Genet* **18:** 3-4.

Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. (1998) New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282(5389)**:682-9.

Collins, J.E., Cole, C.G., Smink, L.J., Garrett, C.L., Leversha, M.A., Soderlund, C.A., Maslen, G.L., Everett, L.A., Rice, K.M., Coffey, A.J., Gregory, S.G., Gwilliam, R., Dunham, A., Davies, A.F., Hassock, S., Todd, C.M., Lehrach, H., Hulsebos, T.J.M., Weissenbach, J., Morrow, B., Kucherlapati, R.S., Wadey, R., Scambler, P.J., Kim, U-J., Simon, M.I., Peyrard, M., Xie, Y-G., Carter, N.P., Durbin, R., Dumanski, J.P., Bentley, D.R. and Dunham, I. (1995) A high-density YAC contig map of human chromosome 22. *Nature* **377(6547 Suppl)**:367-79.

Correa-Villasenor, A., Ferencz, C., Boughman, J.A., and Neill, C.A. (1991) Total anomalous pulmonary venous return: familial and environmental factors. The Baltimore-Washington Infant Study Group. *Teratology* **44(4)**:415-28.

Courtay, C., Heisterkamp, N., Siest, G., and Groffen, J. (1994) Expression of multiple gamma-glutamyl transferase genes in man. *Biochem J* **297 (Pt 3)**:503-8.

Csink, A.K., and Henikoff, S. (1998) Something from nothing: the evolution and utility of satellite repeats. *Trends Genet* **14(5)**:200-4.

Cuny, G., Soriano, P., Macaya, G., Bernardi, G. (1981) The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity. *Eur J Biochem* **115(2)**:227-33.

Delattre, O., Zucman, J., Plougastel, B., Desmaze, C., Melot, T., Peter, M., Kovar, H., Joubert, I., de Jong, P., Rouleau, G., Aurius, A. and Thomas, G. (1992) Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. *Nature* **359(6391)**:162-5.

Depinet, T.W., Zackowski, J.L., Earnshaw, W.C., Kaffe, S., Sekhon, G.S., Stallard, R., Sullivan, B.A., Vance, G.H., Van Dyke, D.L., Willard, H.F., Zinn, A.B., and Schwartz, S. (1997) Characterization of neo-centromeres in marker chromosomes lacking detectable alpha-satellite DNA. *Hum Mol Genet* **6(8)**:1195-204.

Dobyns, W.B., Curry, C.J., Hoyme, H.E., Turlington, L., and Ledbetter, D.H. (1991) Clinical and molecular diagnosis of Miller-Dieker syndrome. *Am J Hum Genet* **48(3)**:584-94 .

Dumanski, J.P., Carlbom, E., Collins, V.P., Nordenskjold, M., Emanuel, B.S., Budarf, M.L., McDermid, H.E., Wolff, R., O'Connell, P., White, R., Lalouel, J-M., and Leppert, M. (1991) A map of 22 loci on human chromosome 22. *Genomics* **11(3)**:709-19.

Edery, P., Lyonnet, S., Mulligan, L.M., Pelet, A., Dow, E., Abel, L., Holder, S., Nihoul-Fekete, C., Ponder, B.A., and Munnich, A. (1994) Mutations of the RET proto-oncogene in Hirschsprung's disease. *Nature* **367(6461)**:378-80. .

Eichler, E.E. (1998) Masquerading repeats: paralogous pitfalls of the human genome. *Genome Res* **8(8)**:758-62. Published erratum appears in Genome Res 1998 Oct;8(10):1095

Eichler, E.E., Budarf, M.L., Rocchi, M., Deaven, L.L., Doggett, N.A., Baldini, A., Nelson, D.L., and Mohrenweiser, H.W. (1997) Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity. *Hum Mol Genet* **6(7)**:991-1002.

Eichler, E.E., Lu, F., Shen, Y., Antonacci, R., Jurecic, V., Doggett, N.A., Moyzis, R.K., Baldini, A., Gibbs, R.A., and Nelson, D.L. (1996) Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum Mol Genet* J **5(7)**:899-912.

Eikenboom, J.C., Vink, T., Briet, E., Sixma, J.J., and Reitsma, P.H. (1994) Multiple substitutions in the von Willebrand factor gene that mimic the pseudogene sequence. *Proc Natl Acad sci USA* **91**:2221-2224.

El-Shanti, H., Hulseberg, D., Murray, J.C., and Patil, S.R., (1993) A three generation minute supernumerary ring 22: association with cat-eye syndrome. *Am J Hum Genet* [suppl] **53:A126**.

Ewart, A.K., Morris, C.A., Atkinson, D., Jin, W., Sternes, K., Spallone, P., Stock, A.D., Leppert, M., and Keating, M.T. (1993) Hemizygosity at the elastin locus in a developmental disorder, Williams syndrome. *Nat Genet* **5(1)**:11-6.

Fisher, E., and Scambler, P. (1994) Human haploinsufficiency--one for sorrow, two for joy. *Nat Genet* **7(1)**:5-7.

Footz, T.K., Birren, B., Minoshima, S., Asakawa, S., Shimizu, N., Riazi, M.A., and McDermid, H.E. (1998) The gene for death agonist BID maps to the region of human 22q11.2 duplicated in cat eye syndrome chromosomes and to mouse chromosome 6. *Genomics* **51(3)**:472-5.

Fraccaro, M., Lindsten, J., Ford, C.E., and Iselius, L. (1980) The 11q;22q translocation: a European collaborative analysis of 43 cases. *Hum Genet* **56(1)**:21-51.

Frizzley, J.K., Stephan, M.J., Lamb, A.N., Jonas, P.P., Hinson, R.M,, Moffitt, D.R., Shkolny, D.L., and McDermid, H.E. (1999) Ring 22 duplication/deletion mosaicism: clinical, cytogenetic, and molecular characterisation. *J Med Genet* **36(3)**:237-41.

Gardiner-Garden, M., and Frommer, M. (1987) CpG islands in vertebrate genomes. *J Mol Biol* **196(2)**:261-82.

Gerald, P.S., Davis, C., Say, B., and Wilkins, J., (1968) A novel chromosomal basis for imperforate anus (the "cat's eye" syndrome). *Pediatr Res* **2:**297 (abstr).

Goldmuntz, E., Driscoll, D., Budarf, M.L., Zackai, E.H., McDonald-McGinn, D.M., Biegel, J.A., and Emanuel, B.S. (1993) Microdeletions of chromosomal region 22q11 in patients with congenital conotruncal cardiac defects. *J Med Genet* **30(10)**:807-12.

Gong, W., Emanuel, B.S., Collins, J., Kim, D.H., Wang, Z., Chen, F., Zhang, G., Roe, B., and Budarf, M.L.(1996) A transcription map of the DiGeorge and velo-cardio-facial syndrome minimal critical region on 22q11. *Hum Mol Genet* **5(6)**:789-800.

Goodman, L. (1998) The human genome project aims for 2003. *Genome Res.* **8(10):**997-9.

Haab, O., (1878) Beitrage zu den angeoborenen Fahlern des Auges. *Von Graefe's Arch Opthalmol* **24**: 257-281.

Haaf, T., Warburton, P.E., and Willard, H.F. (1992) Integration of human alpha-satellite DNA into simian chromosomes: centromere protein binding and disruption of normal chromosome segregation. *Cell* **70(4)**:681-96.

Haluska, F.G., Tsujimoto, Y., and Croce, C.M. (1987) Oncogene activation by chromosome translocation in human malignancy. *Annu Rev Genet* **21**:321-45.

Hattori, M., Adachi, H., Tsujimoto, M., Arai, H., and Inoue, K. (1994) Miller-Dieker lissencephaly gene encodes a subunit of brain platelet-activating factor acetylhydrolase. **Nature 370(6486)**:216-8 [published erratum appears in Nature 1994 Aug 4;370(6488):391].

Homma, K., Matsushita, T., and Natori, S. (1996) Purification, characterization, and cDNA cloning of a novel growth factor from the conditioned medium of NIH-Sape-4, an embryonic cell line of Sarcophaga peregrina (flesh fly). *J Biol Chem* **271(23)**:13770-5.

Hulsebos, T.J., Bijleveld, E.H., Riegman, P.H., Smink, L.J., and Dunham, I. (1996) Identification and characterization of NF1-related loci on human chromosomes 22, 14 and 2. *Hum Genet* **98(1)**:7-11.

Jackson, M.S., Rocchi, M., Thompson, G., Hearn, T., Crosier, M., Guy, J., Kirk, D., Mulligan, L., Ricco, A., Piccininni, S., Marzella, R., Viggiano, L., and Archidiacono, N. (1999) Sequences flanking the centromere of human chromosome 10 are a complex patchwork of arm-specific sequences, stable duplications and unstable sequences with homologies to telomeric and other centromeric locations. *Hum Mol Genet* **8(2)**:205-15.

John, R.M., Robbins, C.A., and Myers, R.M. (1994) Identification of genes within CpG-enriched DNA from human chromosome 4p16.3. *Hum Mol Genet* **3(9)**:1611-6.

Johnson, A., Minoshima, S., Asakawa, S., Shimizu, N., Shizuya, H., Roe, B.A., and McDermid, H.E. (1999) A 1.5-Mb contig within the cat eye syndrome critical region at human chromosome 22q11.2. *Genomics* **57(2)**:306-9.

Kaplan, J.C., Aurias, A., Julier, C., Prieur, M., and Szajnert, M.F. (1987) Human chromosome 22. *J Med Genet* **24(2)**:65-78.

Karlin, S., and Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A* **87(6)**:2264-8.

Karlin, S., and Altschul, S.F. (1993) Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc Natl Acad Sci U S A* **90(12)**:5873-7.

Karpen, G.H., and Allshire, R.C. (1997) The case for epigenetic effects on centromere identity and function. *Trends Genet* **13(12)**:489-96.

Kim, U.J., Shizuya, H., Kang, H.L., Choi, S.S., Garrett, C.L., Smink, L.J., Birren, B.W., Korenberg, J.R., Dunham, I., and Simon, M.I. (1996) A bacterial artificial chromosome-based framework contig map of human chromosome 22q. *Proc Natl Acad Sci U S A* **93(13)**:6297-301.

Knoll, J.H., Asamoah, A., Pletcher, B.A., and Wagstaff, J. (1995) Interstitial duplication of proximal 22q: phenotypic overlap with cat eye syndrome. *Am J Med Genet* **55(2)**:221-4.

Krmpotic, E., Rosnick, M.R., and Zollar, L.M. (1971) Genetic counseling. Secondary nondisjunction in partial trisomy 13. *Obstet Gynecol* **37(3)**:381-90.

Kulp, D., Haussler, D., Reese, M.G., and Eeckman, F.H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *Ismb*. **4**:134-42.

Kundu, T.K., and Rao, M.R. (1999) CpG islands in chromatin organization and gene expression. *J Biochem* (Tokyo) **125(2)**:217-22.

Larin, Z., Fricker, M.D., and Tyler-Smith, C. (1994) De novo formation of several features of a centromere following introduction of a Y alphoid YAC into mammalian cells. *Hum Mol Genet* **3(5)**:689-95.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992) CpG islands as gene markers in the human genome. *Genomics* **13(4)**:1095-107.

Levinson, B., Kenwrick, S., Gamel, P., Fisher, K., and Gitschier, J. (1992) Evidence for a third transcript from the human factor VIII gene. *Genomics* **14(3)**:585-9.

Lotscher, E., Grzeschik, K.-H., Bauer, H. G., Pohlenz, H.-D., Straubinger, B., Zachau, H. G. (1986) Dispersed human immunoglobulin kappa light-chain genes. *Nature* **320**: 456-458.

Lovett, M., Kere, J., and Hinton, L.M. (1991) Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc Natl Acad Sci U S A* **88(21)**:9628-32.

Lupski, J.R., Wise, C.A., Kuwano, A., Pentao, L., Parke, J.T., Glaze, D.G., Ledbetter, D.H., Greenberg, F., and Patel, P.I. (1992) Gene dosage is a mechanism for Charcot-Marie-Tooth disease type 1A. *Nat Genet* **1(1)**:29-33.

Mancuso, D.J., Tuley, E.A., Westfield, L.A., Lester-Mancuso, T.L., Le Beau, M.M., Sorace, J.M., and Sadler, J.E. (1991) Human von Willebrand factor gene and pseudogene: structural analysis and differentiation by polymerase chain reaction. *Biochemistry* **30(1)**:253-69.

May WA, Lessnick, S.L,, Braun, B.S., Klemsz, M., Lewis, B.C., Lunsford, L.B., Hromas, R., and Denny, C.T. (1993) The Ewing's sarcoma EWS/FLI-1 fusion gene encodes a more potent transcriptional activator and is a more powerful transforming gene than FLI-1. *Mol Cell Biol* **13(12)**:7393-8.

McDermid, H.E., Duncan, A.M., Brasch, K.R., Holden, J.J., Magenis, E., Sheehy, R., Burn, J., Kardon, N., Noel, B., Schinzel, A., Teshima, I., and White, B.N. (1986) Characterization of the supernumerary chromosome in cat eye syndrome. *Science* **232(4750)**:646-8.

McDermid, H.E., McTaggart, K.E., Riazi, M.A., Hudson, T.J., Budarf, M.L., Emanuel, B.S., and Bell, C.J. (1996) Long-range mapping and construction of a YAC contig within the cat eye syndrome critical region. *Genome Res* **6(12)**:1149-59.

McDermid, H.E., McTaggart, K.E., Riazi, M.A., Hudson, T.J., Budarf, M.L., Emanuel, B.S., and Bell, C.J. (1996) Long-range mapping and construction of a YAC contig within the cat eye syndrome critical region. *Genome Res* **6(12)**:1149-59.

McKusick, V. (1989) The human genome organisation: history, purposes, and membership. *Genomics* **5**: 385-387.

McKusick, V.A. (1998) Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders. Baltimore: Johns Hopkins University Press, (12th edition).

McTaggart, K.E. M.Sc. thesis (1997) Human cat eye syndrome duplication breakpoints. University of Alberta.

McTaggart, K.E., Budarf, M.L., Driscoll, D.A., Emanuel, B.S., Ferreira, P., and McDermid, H.E. (1998) Cat eye syndrome chromosome breakpoint clustering: identification of two intervals also associated with 22q11 deletion syndrome breakpoints. *Cytogenet Cell Genet* **81(3-4)**:222-8.

Mears, A.J., Duncan, A.M., Budarf, M.L., Emanuel, B.S., Sellinger, B., Siegel-Bartelt, J., Greenberg, C.R., and McDermid, H.E. (1994) Molecular characterization of the marker chromosome associated with cat eye syndrome. *Am J Hum Genet* **55(1)**:134-42.

Mears, A.J. Ph.D. thesis (1995) Molecular characterization of cat eye syndrome. University of Alberta.

Mears, A.J., el-Shanti, H., Murray, J.C., McDermid, H.E., and Patil, S.R. (1995) Minute supernumerary ring chromosome 22 associated with cat eye syndrome: further delineation of the critical region. *Am J Hum Genet* **57(3)**:667-73.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., Ding, W., Bell, R., Rosenthal, J., Hussey, C., Tran, T., McClure, M., Frye, C., Hattier, T., Phelps, R., Haugen-Strano, A., Katcher, H., Yakumo, K., Gholami, Z., Shaffer, D., Stone, S., Bayer, S., Wray, C., Bogden, R., Dayananth, P., Ward, J., Tonin, P., Narod, S., Bristow, P.K., Norris, F.H., Helvering, L., Morrison, P., Rosteck, P., Lai, M., Barrett, C., Lewis, C., Neuhausen, S., Cannon-Albright, L., Goldgar, D., Wiseman, R., Kamb, A., and Skolnick, M.H. (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266(5182)**:66-71.

Minoshima, S., Asakawa, S., Kawasaki, K., Kudoh, J., Shibuya, K., Shintani, A., Aoki, N., Tochigi, J., Johnson, A., Riazi, M.A., McDermid, H.E. Shimizu, Y., and Shimizu, N. (1998) Genomic sequencing and gene identification of the cat eye syndrome chromosome region (22q11.1-q11.2). *Am J Hum Gen* [supp] **63(4)**:1469.

Moore, K.L. and Persaud, T.V.N. (1993) Before we are born: essentials of embryology and birth defects. W.B. Saunders Company, Philadelphia.

Morton, N.E. (1991) Parameters of the human genome. *Proc Natl Acad Sci U S A* **88(17)**:7474-6.

Nesslinger, N. Ph.D. thesis (1994) Characterization of chromosome 22q13.3 deletions. University of Alberta

Ohashi, H., Wakui, K., Seki, K., Niikawa, N., and Fukushima, Y., (1993) Partial cat eye snydrome and supernumerary small ring chromosome 22 detected by microdissection-chromosome painting method. *Am J Hum Genet* [suppl] **53:A586**.

Online Mendelian Inheritance in Man, OMIM (TM). Center for Medical Genetics, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information,National Library of Medicine (Bethesda, MD), 1999. World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/

Parimoo, S., Patanjali, S.R., Shukla, H., Chaplin, D.D., and Weissman, S.M. (1991) cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc Natl Acad Sci U S A* **88(21)**:9623-7.

Patel, P.I., Roa, B.B., Welcher, A.A., Schoener-Scott, R., Trask, B.J., Pentao, L., Snipes, G.J., Garcia, C.A., Francke, U., Shooter, E.M., Lupski, J.R. and Suter, U. (1992) The gene for the peripheral myelin protein PMP-22 is a candidate for Carcot-Marie-Tooth disease type 1A. **Nat Genet 1(3)**:159-65.

Pentao, L., Wise, C.A., Chinault, A.C., Patel, P.I., and Lupski, J.R. (1992) Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. *Nat Genet* **2(4)**:292-300.

Petrij, F., Giles, R. H., Dauwerse, H. G., Saris, J. J., Hennekam, R. C. M., Masuno, M., Tommerup, N., van Ommen, G.-J. B., Goodman, R. H., Peters, D. J. M., and Breuning, M. H. (1995) Rubinstein-Taybi syndrome caused by mutations in the transcriptional co-activator CBP. *Nature* **376**: 348-351.

Pfeiffer, R.A., Heimann, K., and Heiming, E. (1970) Extra chromosome in "cat eye" syndrome. *Lancet* **1(7663)**:97.

Pizzuti, A., Novelli, G., Ratti, A., Amati, F., Bordoni, R., Mandich, P., Bellone, E., Conti, E., Bengala, M., Mari, A., Silani, V., and Dallapiccola, B. (1999) Isolation and characterization of a novel transcript embedded within HIRA, a gene deleted in DiGeorge syndrome. *Mol Genet Metab* **67(3)**:227-35.

Potier, M., Dutriaux, A., Orti, R., Groet, J., Gibelin, N., Karadima, G., Lutfalla, G., Lynn, A., Van Broeckhoven, C., Chakravarti, A., Petersen, M., Nizetic, D., Delabar, J., and Rossier, J. (1998) Two sequence-ready contigs spanning the two copies of a 200-kb duplication on human 21q: partial sequence and polymorphisms. *Genomics* **51(3)**:417-26.

Puech, A., Saint-Jore, B., Funke, B., Gilbert, D.J,. Sirotkin, H., Copeland, N.G., Jenkins, N.A., Kucherlapati, R., Morrow, B., Skoultchi, A.I. (1997) Comparative mapping of the human 22q11 chromosomal region and the orthologous region in mice reveals complex changes in gene organization. *Proc Natl Acad Sci U S A* **94(26)**:14608-13.

Regnier, V., Meddeb, M., Lecointre, G., Richard, F., Duverger, A., Nguyen, V.C., Dutrillaux, B., Bernheim, A., and Danglot, G. (1997) Emergence and scattering of multiple neurofibromatosis (NF1)-related sequences during hominoid evolution suggest a process of pericentromeric interchromosomal transposition. *Hum Mol Genet* **6(1)**:9-16.

Reiner, O., Carrozzo, R., Shen, Y., Wehnert, M., Faustinella, F., Dobyns, W.B., Caskey, C.T., and Ledbetter, D.H. (1993) Isolation of a Miller-Dieker lissencephaly gene containing G protein beta-subunit-like repeats. *Nature* **364(6439)**:717-21.

Reiss, J.A., Weleber, R.G., Brown, M.G., Bangs, C.D., Lovrien, E.W., and Magenis, R.E. (1985) Tandem duplication of proximal 22q: a cause of cat-eye syndrome. *Am J Med Genet* **20(1)**:165-71.

Riazi, M.A. Ph.D. thesis (1998) Transcriptional mapping in the proximal region of human chromosome 22. University of Alberta.

Ritchie, R.J., Mattei, M.G., and Lalande, M. (1998) A large polymorphic repeat in the pericentromeric region of human chromosome 15q contains three partial gene duplications. *Hum Mol Genet* **7(8)**:1253-60.

Romeo, G., Ronchetto, P., Luo, Y., Barone, V., Seri, M., Ceccherini, I., Pasini, B., Bocciardi, R., Lerone, M., and Kaariainen, H.,(1994) Point mutations affecting the tyrosine kinase domain of the RET proto-oncogene in Hirschsprung's disease. *Nature* **367(6461)**:377-8.

Rommens, J.M., Iannuzzi, M.C., Kerem, B., Drumm, M.L., Melmer, G., Dean, M., Rozmahel, R., Cole, J.L., Kennedy, D., Hidaka, N., Zsiga, M., Buchwald, M., Riordan, J.R., Tsui, L-C., Collins, F.S. (1989) Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245(4922)**:1059-65.

Rowley, J.D. (1973) A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243(5405)**:290-3.

Rubinstein, J. H. and Taybi, H. (1963) Broad thumbs and toes and facial abnormalities. *Am. J. Dis. Child.* **105**: 588-608.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular cloning: A laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press.

Sargent, C.A., Dunham, I., and Campbell, R.D. (1989) Identification of multiple HTF-island associated genes in the human major histocompatibility complex class III region. *EMBO J* **8(8)**:2305-12.

Sauer, F., and Jackle, H. (1993) Dimerization and the control of transcription by Kruppel. **Nature 364(6436)**:454-7.

Schachenmann, G., Scmid, W., Fraccaro, M., Mannini, A., Tiepolo, L., Perona, G.P., and Sartori, E., (1965) Chromosomes in coloboma and anal atresia. *Lancet* **2:290**.

Schedl,A., Ross, A., Lee, M., Engelkamp, D., Rashbass, P., van Heyningen, V., and Hastie, N.D. (1996) Influence of PAX6 gene dosage on development: overexpression causes severe eye abnormalities. *Cell* **86(1)**:71-82.

Schinzel, A., Schmid, W., Fraccaro, M., Tiepolo, L., Zuffardi, O., Opitz, J.M., Lindsten, J., Zetterqvist, P., Enell, H., Baccichetti, C., Tenconi, R., and Pagon, R.A. (1981) The "cat eye syndrome": dicentric small marker chromosome probably derived from a no.22 (tetrasomy 22pter to q11) associated with a characteristic phenotype. Report of 11 patients and delineation of the clinical picture. *Hum Genet* **57(2)**:148-58.

Schmickel, R.D., and Knoller, M. (1977) Characterization and localization of the human genes for ribosomal ribonucleic acid. *Pediatr Res* **11(8)**:929-35.

Sentis, C., Ludena, P., and Fernandez-Piqueras, J. (1993) Non-uniform distribution of methylatable CCGG sequences on human chromosomes as shown by in situ methylation. *Chromosoma* **102(4)**:267-71.

Solovyev, V.V.,Salamov, A.A., and Lawrence, C.B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucl.Acids Res.* **22(24)**: 5156-5163.

Sossin, W.S., Kreiner, T., Barinaga, M., Schilling, J., and Scheller, R.H. (1989) A dense core vesicle protein is restricted to the cortex of granules in the exocrine atrial gland of Aplysia california. *J Biol Chem* **264(28)**:16933-40.

Strachan, T., and Read, A.P. (1996) Human molecular genetics. John Wiley & sons, Inc., Publication. New York.

Tazi, J., and Bird, A. (1990) Alternative chromatin structure at CpG islands. *Cell* **60(6)**:909-20.

The Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**:971-983.

The Sanger Centre and the Washington University Genome Sequencing Center (1998) Toward a complete human genome sequence. *Genome Res* **8(11)**:1097-108.

Thompson, M.W., McInnes, R.R., and Willard, H.F. (1991) Genetics in medicine. W.B.saunders Company, Philadelphia.

Ton, C.C., Hirvonen, H., Miwa, H., Weil, M.M., Monaghan, P., Jordan, T., van Heyningen, V., Hastie, N.D., Meijers-Heijboer, H., Drechsler, M., Royer-Pokora, B., Collins, F., Swaroop, A., Strong, L.C., and Saunders, G.F. (1991) Positional cloning and characterization of a paired box- and homeobox-containing gene from the aniridia region. *Cell* **67(6)**:1059-74.

Toniolo, D., D'Urso, M., Martini, G., Persico, M., Tufano, V., Battistuzzi, G., and Luzzatto, L. (1984) Specific methylation pattern at the 3' end of the human housekeeping gene for glucose 6-phosphate dehydrogenase. *EMBO J* **3(9)**:1987-95.

Toomey, K.E., Mohandas, T., Leisti, J., Szalay, G., and Kaback, M.M. (1977) Further delineation of the supernumerary chromosome in the Cat-Eye syndrome. *Clin Genet* **12(5)**:275-84.

Tribioli, C., Tamanini, F., Patrosso, C., Milanesi, L., Villa, A., Pergolizzi, R., Maestrini, E., Rivella, S., Bione, S., Mancini, M.,Vezzoni, P., and Tonilo, D. (1992) Methylation and sequence analysis around EagI sites: identification of 28 new CpG islands in XQ24-XQ28. *Nucleic Acids Res* **20(4)**:727-33.

Trofatter, J.A., Long, K.R., Murrell, J.R., Stotler, C.J., Gusella, J.F., and Buckler, A.J. (1995) An expression-independent catalog of genes from human chromosome 22. *Genome Res* **5(3)**:214-24.

Trofatter, J.A., MacCollin, M.M., Rutter, J.L., Murrell, J.R., Duyao, M.P., Parry, D.M., Eldridge, R., Kley, N., Menon, A.G., Pulaski, K., Haase, V.H., Ambrose, C.M., Munroe, D., Bove, C., Haines, J.L., Martuza, R.L., MacDonald, M.E., Seizinger, B.R., Short, M.P., Buckler, A.J., and Gusella, J.F. (1993) A novel moesin-, ezrin-, radixin-like gene is a candidate for the neurofibromatosis 2 tumor suppressor. *Cell* **72**:791-800. [erratum appears in *Cell* **75(4)**:826].

Tyler-Smith, C., Oakey, R.J., Larin, Z., Fisher, R.B., Crocker, M., Affara, N.A., Ferguson-Smith, M.A., Muenke, M., Zuffardi, O., and Jobling, M.A. (1993) Localization of DNA sequences required for human centromere function through an analysis of rearranged Y chromosomes. *Nat Genet* **5(4)**:368-75.

Valdes, J.M., Tagle, D.A., and Collins, F.S.(1994) Island rescue PCR: a rapid and efficient method for isolating transcribed sequences from yeast artificial chromosomes and cosmids. *Proc Natl Acad Sci U S A* **91(12)**:5377-81.

Viskochil, D., Cawthon, R., O'Connell, P., Xu, G.F., Stevens, J., Culver, M., Carey, J., and White, R. (1991) The gene encoding the oligodendrocyte-myelin glycoprotein is embedded within the neurofibromatosis type 1 gene. *Mol Cell Biol* **11(2)**:906-12.

Vortkamp, A., Franz, T., Gessler, M., and Grzeschik, K.H. (1992) Deletion of GLI3 supports the homology of the human Greig cephalopolysyndactyly syndrome (GCPS) and the mouse mutant extra toes (Xt). *Mamm Genome* **3(8)**:461-3.

Vortkamp, A., Gessler, M., and Grzeschik, K.H. (1991) GLI3 zinc-finger gene interrupted by translocations in Greig syndrome families. *Nature* **352(6335)**:539-40.

Williams, B.C., Murphy, T.D., Goldberg, M.L., and Karpen, G.H. (1998) Neocentromere activity of structurally acentric mini-chromosomes in Drosophila. *Nat Genet* **18(1)**:30-7.

Wong, Ph.D. thesis (1998) Transcriptional mapping in 22q microdeletion. University of Alberta

Wong, Z., Royle, N.J., and Jeffreys, A.J. (1990) A novel human DNA polymorphism resulting from transfer of DNA from chromosome 6 to chromosome 16. *Genomics* **7(2)**:222-34.

Worley, K.C., Wiese, B.A., and Smith, R.F. (1995) BEAUTY: an enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. *Genome Res* **5(2)**:173-84.

Yao, Z., Spriggs, M.K., Derry, J.M., Strockbine, L., Park, L.S., VandenBos, T., Zappone, J.D., Painter, S.L., and Armitage, R.J. (1997) Molecular characterization of the human interleukin (IL)-17 receptor. *Cytokine* **9(11)**:794-800.

Zackai, E.H., and Emanuel, B.S. (1980) Site-specific reciprocal translocation, t(11;22) (q23;q11), in several unrelated families with 3:1 meiotic disjunction. *Am J Med Genet* **7(4)**:507-21.

Zhang, M.Q. (1997) Identification of Protein Coding Regions in the Human Genome Based on Quadratic Discriminant Analysis. *PNAS* **94**:565-568.